nature genetics

Supplementary information

https://doi.org/10.1038/s41588-024-01933-1

Multivariate genomic analysis of 5 million people elucidates the genetic architecture of shared components of the metabolic syndrome

In the format provided by the authors and unedited

Contents

Supplementary Note

1. Summary of the clinical definitions of metabolic syndrome

Metabolic syndrome (MetS) is a collection of risk factors that increase the risk of cardiovascular disease and type 2 diabetes (T2D). Despite its seemingly straightforward definition, it is still a challenge to diagnose MetS clinically[1-3].

The initial definition of MetS was established in 1998 by the World Health Organization[4], highlighting insulin resistance along with any two of the following criteria: obesity, hypertension, high triglyceride (TG) levels, reduced high-density lipoprotein cholesterol (HDL-C) levels, or microalbuminuria. In 2001, the National Cholesterol Education Program Adult Treatment Panel III (ATP III)[5] revised the criteria, neglecting insulin resistance as a mandatory requirement and requiring three out of five specific conditions for MetS diagnosis: abdominal obesity, elevated TG, reduced HDL-C, elevated BP, and elevated fasting glucose (FG). However, the American Association of Clinical Endocrinologists (AACE)[6] once again underscored the importance of insulin resistance in 2003.

By 2005, the International Diabetes Federation (IDF)[7] and the American Heart Association/National Heart, Lung, and Blood Institute (AHA/NHLBI)[8] had proposed definitions similar to ATP III, with the IDF emphasizing the centrality of abdominal obesity and the AHA/NHLBI deeming it less critical. Eventually, a consensus was reached on the following criteria: elevated waist circumference (WC), elevated TG, reduced HDL-C, elevated BP, and elevated FG, with abdominal obesity not being a mandatory factor for MetS diagnosis but requiring different WC thresholds based on sex and ethnicity.

Similarly, the clinical definition of MetS undergoes periodic revisions and updates. The

ongoing progression of MetS can be attributed to several factors. As research on MetS advances, a new understanding of its underlying mechanisms and pathophysiology is continuously being incorporated, leading various organizations to emphasize different aspects of MetS. Additionally, the prevalence and significance of MetS components can differ among populations and ethnicities, necessitating modifications to the diagnostic criteria to reflect these variations. The challenge of reaching an agreement among organizations, each with its own set of perspectives and priorities, leads to a diversity of criteria for defining MetS.

In this study, we adopted the definition of MetS outlined by the IDF and AHA/NHLBI, which specifies five criteria. However, we allowed for certain flexibility by including additional traits and diseases closely linked to these criteria, such as body mass index and T2D. The inclusion of additional genetically correlated traits can be advantageous for this study, as it uses various traits to conduct multivariate Genome-wide Association Studies (GWAS). This approach enhances the ability to identify new genetic associations by boosting statistical power and improving the accuracy of polygenic risk scores, as measurement of one trait can offer information on the genetic values of other related traits[9-11]. While MetS is recognized as a risk factor for T2D, T2D itself can be viewed as a component of MetS, especially because the AHA/NHLBI threshold for fasting glucose (≥100 mg/dL) encompasses most patients with T2D. The overlap of common risk factors between T2D and other MetS components, such as obesity, dyslipidemia, and hypertension, highlights the correlated nature of these conditions. This underscores the importance of identifying individuals at risk of both MetS and T2D to effectively implement early intervention and prevention strategies.

Summary of the metabolic syndrome diagnosis criteria

	Organization	WHO[4]	ATP III[5]	AACE[6]	IDF[7]	AHA/NHLBI[8]
	Year	1998	2001	2003	2005	2005
Category	**MetS diagnosis criteria**	Mandate insulin resistance and ≥2 of the following	≥3 of the following	Mandate insulin resistance and any of the following	Mandate obesity and ≥2 of the following	≥3 of the following
Obesity	BMI	>30 kg/m^2 (either BMI or WHR)	-	≥25 kg/m^2	≥30 kg/m^2 (if stratifies, WC is not required)	-
Obesity	WC	-	Sex-specific; ≥102 cm in male, ≥80 cm in female	-	Population- and sex-specific; ≥94 cm in EUR male, ≥80 cm in EUR female	Sex-specific; ≥102 cm in male, ≥88 cm in female
Obesity	WHR	Sex-specific; >0.9 in male, >0.85 in female	-	-	-	-
Hypertension	BP	Systolic ≥140 mm Hg and/or diastolic ≥90 mm Hg	Systolic ≥130 mm Hg and/or diastolic ≥85 mm Hg; Hypertension prescription	Systolic ≥130 mm Hg and/or diastolic ≥85 mm Hg	Systolic ≥130 mm Hg and/or diastolic ≥85 mm Hg; Hypertension prescription	Systolic ≥130 mm Hg and/or diastolic ≥85 mm Hg; Treatment with antihypertensive drugs with a history of hypertension
Insulin resistance	DM	Diagnosed with T2D	FG criterion does not exclude diabetes[8]	-	Diagnosed with T2D	Majority of patients with T2D included based on the FG criterion[1]
Insulin resistance	FG	IGT or IFG	>110 mg/dL	IGT or IFG without diabetes	≥100 mg/dL	≥100 mg/dL
Dyslipidemia	TG	≥150 mg/dL	≥150 mg/dL	≥150 mg/dL	≥150 mg/dL	≥150 mg/dL
Dyslipidemia	HDL	Sex-specific; <35 mg/dL in male, <39 mg/dL in female	Sex-specific; <40 mg/dL in male, <50 mg/dL in female	Sex-specific; <40 mg/dL in male, <50 mg/dL in female	Sex-specific; <40 mg/dL in male, <50 mg/dL in female	Sex-specific; <40 mg/dL in male, <50 mg/dL in female
Other	Other	Microalbuminuria	-	Family history of T2D	-	-

Abbreviation: WHO, World Health Organization; ATP III, National Cholesterol Education Program Adult Treatment Panel III; AACE, American Association of Clinical Endocrinologists; IDF, International Diabetes Federation; AHA/NHLBI, American Heart Association/National Heart, Lung, and Blood Institute; MetS, metabolic syndrome; BMI, body mass index; WC, waist circumference; EUR, European; WHR, waist-to-hip ratio; BP, blood pressure; DM, diabetes mellitus; T2D, type 2 diabetes; FG, fasting glucose; IGT, impaired glucose intolerance; IFG, impaired fasting glucose; TG, triglycerides; HDL, high-density lipoprotein cholesterol

2. Functional annotation

We performed functional annotation of metabolic syndrome (MetS) genetic signal from a multivariate genome-wide association study (GWAS) by using ANNOVAR[12] implemented in FUMA[13]. The independent significant single-nucleotide polymorphisms (SNPs) and their SNPs in linkage disequilibrium (LD) were annotated with relevant biological functions. The annotation leveraging ANNOVAR is performed by matching the SNP's chromosome, base position, reference allele, and alternative allele to the databases.

The databases used for the functional annotation in this study were ANNOVAR categories[12], CADD scores[14], Regulome DB scores[15], and chromatin states[16,17]. A detailed description of each database can be found from corresponding studies, as well as Jansen *et al.*[18], and a summarized description can be found below:

1) ANNOVAR categories are divided into exonic, intergenic, 5′/3′-UTR, splicing site, and upstream/downstream, where the SNPs are annotated by their genic position.
2) The CADD scores are the predicted measure of how deleterious a SNP is for a protein function and higher scores represent higher deleteriousness. We used the default threshold of CADD score > 12.37 to define a SNP as being deleterious.
3) The RegulomeDB scores are categorized from 1a to 7 based on eQTLs and chromatin marks, where 1a indicates that SNP is most likely to have a regulatory function. The categories are as follows: 1a: eQTL + transcription factor (TF) binding + matched TF motif + matched DNase footprint + DNase peak, 1b: eQTL + TF binding + any motif + DNase footprint + DNase peak, 1c: eQTL + TF binding + matched TF motif + DNase peak, 1d: eQTL + TF binding + any motif + DNase peak, 1e: eQTL + TF binding + matched TF motif, 1f: eQTL + TF binding/DNase peak, 2a: TF binding + matched TF motif + matched DNase footprint + DNase peak, 2b: TF binding any motif + DNase footprint + DNase peak, 2c: TF binding + matched TF motif + DNase peak,

3a: TF binding + any motif + DNase peak, 3b: TF binding + matched TF motif, 4: TF binding + DNase peak, 5: TF binding or DNase peak, 6: motif hit, and 7: other. The chromatin states represent the accessibility of the genomic region for every 200 bp using 15 categorical states predicted by ChromHMM[17] using 5 chromatin marks for 127 epigenomes.

4) The chromatin states from 1 to 7 represent open chromatin states and the lower chromatin state indicates higher accessibility. The 15 chromatin states are as follows: 1: active transcription start site (TSS), 2: flanking active TSS, 3: transcription at gene 5′ and 3′, 4: strong transcription, 5: weak transcription, 6: genic enhancers, 7: enhancers, 8: zinc-finger protein and repeats, 9: heterochromatin, 10: bivalence/poised TSS, 11: flanking bivalent TSS/enhancer, 12: bivalent enhancer, 13: repressed polycomb, 14: weak repressed polycomb, and 15: quiescent/low.

Most of COJO MetS SNPs were intronic (*n* SNP = 588, 45%) or intergenic (*n* SNP = 427, 32.7%), yet 26 COJO SNPs (2%) were nonsynonymous. Based on the CADD score, an SNP with a CADD score greater than 12.35 is considered deleterious. We observed 112 COJO MetS SNPs being deleterious in which the SNP with the highest probability of exerting deleterious protein effect was rs3764002 (CADD score = 28.1). The RegulomeDB scores categorized 44 COJO SNPs under RegulomeDB score 1 (i.e., a high likelihood of having a regulatory function) based on potential regulatory function. There were 365 COJO SNPs annotated under open chromatin states where 19 COJO SNPs were in the active transcription start site (TSS) with the highest accessibility (**Supplementary Fig. 4** and **Supplementary Table 13**).

3. SNP to gene mapping using FUMA and MAGMA gene-based analysis

FUMA was leveraged to map independent significant SNPs of MetS GWAS and their SNPs in

LD to relevant genes. The gene mapping was performed by three different strategies available in FUMA, namely positional mapping, expression quantitative trait loci (eQTL) mapping, and chromatin interaction mapping. For all three mapping strategies, we additionally filtered the SNPs based on the functional annotations as follows: minimum CADD score ≥12.37, maximum Regulome DB score of 7, all tissue- and/or cell-types available for 15-core chromatin state, and maximum 15-core chromatin state of 7.

A brief description of three gene mapping strategies is as follows:

1) Positional mapping maps SNPs to protein-coding genes with the physical distance between the SNP and the gene with a default window of 10 kb.
2) eQTL mapping maps independent significant SNPs and SNPs in LD to genes where these SNPs are likely to affect the expression of those genes within 1 Mb. We used a Benjamini-Hochberg FDR of 0.05 to define a significant eQTL association. All tissue types from data repositories available in FUMA for eQTL mapping which include EyeGEx[19], eQTL catalogue[20], PsychENCODE[21], van der Wijst *et al.* scRNA eQTLs[22], DICE[23], eQTLGen[24], Blood eQTLs[25], MuTHER[26], xQTL server[27], CommInMind Consortium[28], BRAINEAC[29], and GTEx v8[30].
3) Chromatin interaction mapping maps independent significant SNPs and SNPs in LD to a gene if there is three-dimensional DNA-DNA interaction between a region containing the SNP and another region that overlaps with the promoter region of the gene (250 bp upstream and 500 bp downstream of the TSS). The interacting region may span multiple genes as chromatin interactions are often defined with a certain resolution (e.g., 40 kb). Due to the absence of distance boundaries, chromatin interaction mapping can involve long-range interactions. We used an FDR P-value of 1×10^{-6} to define significant interactions. The data repositories used for chromatin interactions are 21 tissue and cell

types of Hi-C data[31], adult and fetal cortex Hi-C data[32], FANTOM[33], and PsychENCODE[21].

The independent SNPs from MetS GWAS and SNPs in LD were mapped to genes using three gene mapping strategies. The positional mapping, eQTL mapping, and chromatin interaction mapping identified 885, 3,008, and 4,660 genes, respectively, where 505 genes were mapped by all three strategies (**Supplementary Table 14**). We also conducted the gene-based analysis using MAGMA[34] which aggregates the association *P*-values of all SNPs physically located in a gene (**Methods**). MAGMA gene-based analysis identified 2,092 genes after Bonferroni correction (*P*-value $<2.83 \times 10^{-6}$) (**Supplementary Table 15**). In total, 414 genes were mapped by all three gene mapping strategies in FUMA and MAGMA gene-based analysis (**Supplementary Fig. 5**).

4. Heterogeneity test of Q_{SNP}

We conducted a heterogeneity test for the MetS factor using genomic SEM (**Supplementary Fig. 8**). This allows us to identify the SNPs having a pleiotropic effect that operates through the MetS factor from the SNPs having a heterogenous effect on one or more first-order factors (i.e., obesity factor, insulin resistance/hypertension factor, and dyslipidemia factor), which are referred to as Q_{SNP}. We could expect a majority of COJO MetS SNPs identified to be in different regions from the Q_{SNP} signal if the COJO MetS SNPs are true pleiotropic signals operating through the MetS factor[35]. Thus, we compared the independence between COJO SNPs of MetS GWAS and lead SNPs of MetS Q_{SNP}. In total, 1,203 Q_{SNP} lead SNPs were identified using an identical clumping approach, and 751 of 1,307 (57.5%) GWAS COJO SNPs were independent of Q_{SNP} lead SNPs. Additionally, only 273 of 1,307 (20.9%) COJO MetS SNPs showed genome-wide significance (*P*-value $<5 \times 10^{-8}$) in Q_{SNP}.

5. Cell- and tissue-specific enrichment and MAGMA gene-set and gene-property analyses

We used LDSC regression for specifically expressed genes (LDSE-SEG)[36] to investigate whether the MetS genetic signal is enriched in specific cell or tissue types by gene expression or activating histone marks. Using multi-tissue and cell gene expression data from Franke lab[37], we observed 14 Bonferroni significant (P-value <0.05/205 = 2.44×10^{-4}) enrichment which includes hippocampus (P-value = 1.30×10^{-10}), limbic system (P-value = 1.40×10^{-10}), and brain (P-value = 1.47×10^{-8}). Especially, the enrichment in brain tissue was also validated from multi-tissue chromatin-based annotation data where 24 brain-related tissues were significantly enriched after Bonferroni correction (P-value <0.05/489 = 1.02×10^{-4}), and enrichment in neurons (P-value = 0.013) among three brain cell types (Bonferroni correction at P-value <0.017) was observed as well (**Supplementary Tables 22–24**). We also conducted MAGMA gene-property analysis for gene expression from 54 specific tissues and found eight brain-specific tissue associations with MetS after Bonferroni correction (P-value <0.05/54 = 9.26×10^{-4}), which aligns with the results from LDSC-SEG (**Supplementary Table 25**). The MAGMA gene-property analysis with 11 brain developmental stages further identified enrichment in prenatal stages (**Supplementary Table 26**).

To further investigate the biological and molecular function of the genes, MAGMA gene-set analysis was conducted against 15,481 curated and gene ontology (GO) gene sets from the Molecular Signature Database (MsigDB v7)[38]. In total, 25 gene sets were significantly associated with MetS and these were related to molecular functions, lipid-relevant functions, and neuronal functions (**Supplementary Table 29**).

6. Summary of 11 MetS genes prioritized using Summary-based Mendelian randomization

A summary of the 11 genes associated with MetS is described below. This information was queried from GeneCards[39], the Open Targets Platform[40], and the International Mouse Phenotyping Consortium (IMPC)[41] (accessed on 2024.04.01).

AMHR2

The *AMHR2* (Anti-Mullerian Hormone Receptor Type 2) encodes a receptor that interacts with the anti-Mullerian hormone (AMH), playing a crucial role alongside testosterone in the differentiation of male sexual characteristics. AMH and testosterone are synthesized in the testes; they are produced by distinct cell types and exert divergent biological effects. While testosterone facilitates the formation of male genital structures, AMH binding to this receptor inhibits the development of Müllerian ducts into the uterus and fallopian tubes. Although it is highly associated with persistent Müllerian duct syndrome, it is associated with neurodegenerative diseases, vitamin D levels, and body mass index (BMI).

BCL7B

The *BCL7B* (BAF Chromatin Remodeling Complex Subunit BCL7B) belongs to the BCL7 family, which comprises BCL7A, BCL7B, and BCL7C proteins, with this specific gene encoding the BCL7B protein. Moreover, it acts as a positive regulator of apoptosis. This gene is evolutionarily conserved from *C. elegans* to humans, highlighting its biological importance across species. Moreover, its GO annotation highlighted its role in actin binding. The metabolic traits associated with this gene include TG levels, BMI-adjusted waist-hip ratio, and diastolic blood pressure. IMPC showed a significant association with increased grip strength, abnormal retinal inner nuclear layer morphology, preweaning lethality (incomplete penetrance), and

decreased total retinal thickness. *BCL7B* was pinpointed as a pleiotropic gene that contributes to both MetS and inflammation and is associated with TG, high-density lipoprotein cholesterol (HDL), and C-reactive protein (CRP)[42].

FEZ2

The *FEZ2* (Fasciculation and elongation protein Zeta 2) plays a crucial role in normal axonal bundling and elongation of axon bundles. Moreover, it is associated with metabolic traits, such as body fat percentage, lean body mass, BMI, and HDL levels.

HM13

The *HM13* (Histocompatibility Minor 13) encodes a protein localized to the endoplasmic reticulum, pivotal for the intramembrane proteolysis of signal peptides post-cleavage from a precursor protein, a process vital for producing human lymphocyte antigen-E epitopes for immune recognition and processing the hepatitis C virus core protein. Characterized by motifs typical of presenilin-type aspartic proteases, this protein is integral to the membrane, and research has uncovered multiple isoforms through various transcript variants. Moreover, it plays a role in the cellular response to stimuli and cytoprotection mediated by *HMOX1*. From IMPC, *HM13* showed a significant association with abnormal snout morphology, increased bone mineral content, preweaning lethality (complete penetrance), and decreased fasting circulating glucose levels. Moreover, it is associated with diastolic blood pressure, BMI, and HDL and TG levels. *HM13* has been suggested as a potential therapeutic target for hepatocellular carcinoma, in which MetS is a major risk factor[43].

MED23

The *MED23* (Mediator Complex Subunit 23) is a key component of the Mediator complex that plays a vital role in the transcriptional activation process by facilitating the interaction between

gene-specific regulatory proteins and the basal RNA polymerase II transcription machinery, and it is involved in various cellular pathways, including those activated by *SP1* and thyroid hormone receptors, and is essential for the efficient transcription of nearly all RNA polymerase II-dependent genes. Additionally, *MED23* has been implicated in disease processes, acting as a metastasis suppressor, and is associated with certain intellectual developmental disorders. From IMPC, *MED23* showed a significant association with preweaning lethality (complete penetrance), decreased bone mineral content, increased red blood cell distribution width, decreased circulating fructosamine levels, increased circulating HDL levels, and increased lean body mass. Metabolic traits that showed an association were BMI, T2D, and fat body mass. In addition, *MED23* may be associated with MetS and cardiovascular diseases owing to its functional involvement in adipocyte and smooth muscle cell differentiation[44].

MLXIPL

The *MLXIPL* (MLX Interacting Protein Like) is a transcription factor belonging to the Myc/Max/Mad superfamily. This factor engages in the formation of a heterodimeric complex (in a glucose-dependent manner), which activates the carbohydrate response element (ChoRE) motifs observed in the promoters of genes involved in triglyceride synthesis. Additionally, it plays a role in pathways such as the integration of energy metabolism and the regulatory pathway of angiopoietin-like protein 8. GO annotations were ascribed to their functions in DNA-binding transcription factor activity and protein heterodimerization activity. Metabolic traits associated with this gene are TG measurement, HDL, low-density lipoprotein, and total cholesterol levels. This gene, like *BCL7B*, is associated with both metabolic traits (TG and HDL) and an inflammation marker (CRP)[42]. Given that *MLXIPL* is linked with the synthesis of TG from excess carbohydrates, which highlights its role in lipid metabolism, MLXIPL appears to be associated with MetS.

MYO1F

Myosin functions as a molecular motor, transforming the energy derived from ATP hydrolysis into mechanical force exerted on actin filaments. *MYO1F* (Myosin IF) encodes a distinct type of myosin implicated in the transport of membrane-bound organelles within cells. The malfunction of this gene has been linked to hearing impairments. Characterized by their ATPase activity, myosins, particularly the unconventional types encoded by this gene, are key to intracellular transport processes. Moreover, it is associated with neurodegenerative diseases, but GWAS shows evidence of an association with BMI-adjusted waist circumference, smoking behavior, HDL measurement, and TG measurement.

RBM6

RBM6 (RNA Binding Motif Protein 6) has RNA-binding capabilities and is expected to play a significant role in mRNA splicing through the spliceosome mechanism. This protein is also expected to function in the nucleus. According to GO annotations, this gene is associated with nucleic acid- and RNA-binding functions, highlighting its critical role in genetic regulation. Regarding IMPC, *RBM6* was significantly associated with decreased fasting circulating glucose levels and grip strength. Moreover, it has been associated with HDL cholesterol measurements, body fat distribution, and T2D.

RFT1

RFT1 (RFT1 homolog) encodes an enzyme crucial for N-glycosylation of proteins, specifically catalyzing the translocation of the Man(5)GlcNAc(2)-PP-Dol intermediate across the endoplasmic reticulum membrane. Mutations in this gene are linked to congenital disorders of glycosylation, and their function is essential in the glycosylation pathway and protein metabolism, with a notable role in lipid transporter activity. According to the IMPC database,

it is significantly associated with decreased erythrocyte cell number, decreased lymphocyte cell number, embryonic lethality before organogenesis, embryonic lethality before the tooth bud stage, increased basophil cell number, increased hemoglobin content, increased mean corpuscular hemoglobin, increased mean corpuscular volume, increased monocyte cell number, and long tibia.

SP1

The *SP1* (Sp1 Transcription Factor) encodes a zinc finger transcription factor that plays a crucial role in various cellular processes such as differentiation, growth, apoptosis, immune response, DNA damage response, and chromatin remodeling, by binding to GC-rich motifs in many promoters, and it acts as both an activator and a repressor of transcription and is significantly influenced by post-translational modifications. They are strongly associated with neurodegenerative diseases. Although a direct relationship between *SP1* and MetS has not yet been established, *SP1* regulates genes associated with hyperinsulinemia, T2D, and MetS in response to insulin[45]. Moreover, it regulates insulin signaling by controlling the transcription of insulin receptors.

STRA13 (CENPX)

STRA13, also known as *CENPX* (Centromere Protein X), is a protein-coding gene involved in DNA binding that plays a crucial role in replication fork processing, resolution of meiotic recombination intermediates, and the Fanconi anemia pathway. The metabolic traits that showed an association with this gene included BMI-adjusted hip circumference, body height, and BMI-adjusted waist-hip ratio. Moreover, it is significantly associated with decreased locomotor activity ($P = 1.88 \times 10^{-11}$), according to IMPC. *STRA13* regulates adipogenesis by influencing the activity of hypoxia-inducible factor 1-alpha (HIF-1α), and it regulates

lipogenesis in the liver by acting as an insulin-responsive gene that is activated via the PI3K pathway and influences the expression of other insulin-sensitive genes. Additionally, *STRA13* expression is induced by glucose through the action of carbohydrate response element-binding protein (ChREBP) at its promoter and participates in a feedback mechanism that inhibits ChREBP's activity to prevent excessive lipogenesis. The dysregulation of *STRA13* may contribute to MetS, highlighting its importance in metabolic regulation[46].

7. Metabolic syndrome phenotyping in UK Biobank

Since MetS phenotype is absent in UK Biobank (UKB) cohort, we assigned the dichotomous MetS status based on AHA/NHLBI[1] that consists of five criteria as follows: population- and sex-specific elevated waist circumference (WC, European with WC_{men} ≥102 cm and WC_{women} ≥88 cm), elevated triglycerides (TG, TG ≥150 mg/dL), sex-specific reduced high-density lipoprotein cholesterol (HDL-C, HDL-C_{men} <40 mg/dL and HDL-C_{women} < 50mg/dL), elevated blood pressure (BP, systolic ≥130 mmHg and diastolic ≥85 mmHg), and elevated fasting glucose (FG, FG ≥100 mg/dL). The individuals who satisfied three out of five criteria based on AHA/NHLBI were categorized as MetS positive. The phenotypic values of WC, TG, HDL-C, systolic blood pressure, and diastolic blood pressure were available in UKB with data field ID of 48, 30870, 30760, 4080, and 4079, respectively. FG was not directly available in UKB; hence we defined the FG based on glucose level (UKB data field ID = 30740) and fasting time (UKB data field ID = 74). The glucose level of UKB individuals with fasting time greater than eight hours was considered as fasting glucose. The total number of UKB individuals considered for MetS dichotomization was 11,139 (4,641 and 6,498 assigned as MetS positive and negative, respectively). It is a relatively small sample size considering the total number of samples in UKB due to a limited number of individuals with fasting time greater than eight hours.

8. Korean Genome and Epidemiology Study (KoGES)

8.1. KoGES description

The KoGES[47] is a large prospective cohort study, initiated by the South Korean government (National Research Institute of Health [NIH], Centers for Disease Control and Prevention and the Ministry of Health and Welfare, South Korea), and consists of three population-based studies as follows: Ansan and Ansung (Ansan/Ansung) study, Health Examinee (HEXA) study, and Cardiovascular Disease Association (CAVAS) study. All individuals in the studies were recruited from the national health examinee registry and the description of each study is as follows: 1) Ansan/Ansung study is a community-based cohort consisting of 10,030 individuals aged 40-69 years old who live in Asan or Ansung city, 2) HEXA study is an urban-based cohort consisting of 173,357 individuals aged 40-79 years old, and 3) CAVAS is a rural-based cohort consisting of 28,338 individuals aged 40-69 years old. The KoGES provides both epidemiology and genetic data from the aforementioned studies where we obtained 5,493, 58,693, and 8,105 individuals from Ansan/Ansung, HEXA, and CAVAS studies, respectively.

8.2. MetS phenotyping in KoGES

We used criteria defined by AHA/NHLBI was used to define MetS for individuals in KoGES. The criteria are as follows: elevated waist circumference (Asian with WC_{men} ≥90 cm and WC_{women} ≥80 cm, elevated TG (TG ≥150 mg/dL), sex-specific reduced HDL-C (HDL-C_{men} <40 mg/dL and HDL-C_{women} <50 mg/dL), elevated blood pressure (systolic ≥130 mmHg and diastolic ≥85 mmHg), and elevated FG (FG ≥100 mg/dL). The individuals were categorized as MetS positive if they satisfied three out of five criteria. The individuals in the CAVAS study were excluded from the analysis as epidemiology data for FG was absent.

8.3. KoGES genotype data quality control and imputation

The KoGES provides genome-wide genotype data from the Korean Biobank Array[48], which is a Korean-specific chip. The genotype data QC was performed using PLINK v1.07, and sample and genotype QC filtering criteria were as follows: heterozygosity rate, exceeding ± 5 s.d.; sex mismatch; missing call rate, <0.02; MAF, <0.01; Hardy-Weinberg equilibrium, $P < 1 \times 10^{-6}$; removal of sex chromosomes. Then, the imputation was conducted using the Michigan Imputation Serve[49] which utilizes Eagle v2.4[50] for phasing and Minimac 4 for genotype imputation with Haplotype Reference Consortium (HRC) r1.1 2016 as reference genome panel. The imputed SNPs were further filtered with genotype quality (R^2 <0.8). We retained 64,100 samples (*n* Ansan/Ansung = 5,487, *n* HEXA = 58,613) and 5,394,481 SNPs as a result of genotype QC and imputation.

9. Polygenic risk score computation using PRSice–2

The PRSice–2[51] employs a pruning and thresholding (P+T) strategy using various subsets of independent SNPs at different *P*-value cutoffs to calculate the polygenic risk score (PRS). First, SNPs underwent a process called clumping (also known as pruning) to identify independent SNPs (using a 500 kb window and an r^2 of 0.1). Subsequently, these independent SNPs were categorized into subsets according to *P*-value thresholds (0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, and 1), and the PRS was calculated for each group. The best *P*-value threshold is determined by identifying the PRS that accounts for the largest variance in the target phenotype (i.e., MetS) within the validation cohort. Although using the same validation data for the best *P*-value selection as the target data for the PRS analysis of incremental R^2 might predispose to overfitting, this potential bias is substantially mitigated when the validation set is sufficiently

large[52].

10. Association between polygenic risk score and cardiovascular disease incidence rate through multivariable Cox regression analysis

MetS is a known risk factor for cardiovascular disease (CVD). We assessed whether the MetS PRS showed better stratification and a stronger association with the CVD incidence rate in the UKB than the PRS of MetS components. We followed the previous work by Yun *et al*.[53] to define CVD incidence and conducted multivariable Cox regression analyses.

Briefly, individuals with ischemic stroke, hemorrhagic stroke, peripheral artery disease, heart failure, or arterial fibrillation/flutter were categorized as having CVD incidence. Collectively, among 352,781 individuals without prior CVD history, 35,711 individuals had CVD incidence with a median follow-up period of 11 years (interquartile range of 10.2–11.7). The overall CVD incidence rate was 9.67 incidence per 1000 persons-year (95% CI = 9.57–9.77). Multivariable Cox regression analysis was used to assess the relationship between the PRS of interest and the incidence rate of CVD. This assessment was conducted by stratifying PRS into distinct groups and considering the PRS itself. The stratification of PRS resulted in four groups: low risk (<20%), intermediate risk (20–80%), high risk (80–99%), and very high risk (>99%). The hazard ratios (HRs) of PRS groups were computed using the low-risk group as a reference. Age, sex, and the first 10 principal genetic components were adjusted as covariates. The analysis was conducted using the R package *survival* v.3.5.7.

The CVD incidence rate for the stratified PRS was evaluated (**Supplementary Fig. 9**). Moreover, it is evident from the visualization that the MetS PRS demonstrates the most pronounced discrimination in CVD incidence compared to the PRS of MetS components,

exhibiting the highest HR of 1.33 (95% CI = 1.28–1.38) in the very high-risk group relative to the low-risk group. Furthermore, the MetS PRS shows the strongest association with CVD incidence, with an HR of 1.11 per unit standard deviation (95% CI = 1.10–1.13, *P*-value = 4.66×10^{-76}) (**Supplementary Table 35**). Collectively, these findings suggest that the MetS PRS is superior in predicting adverse outcomes, such as CVD, compared to the PRS of its individual components.

11. Causal effect estimation using constrained maximum likelihood and model averaging-based (cML-MA) Mendelian randomization (MR)

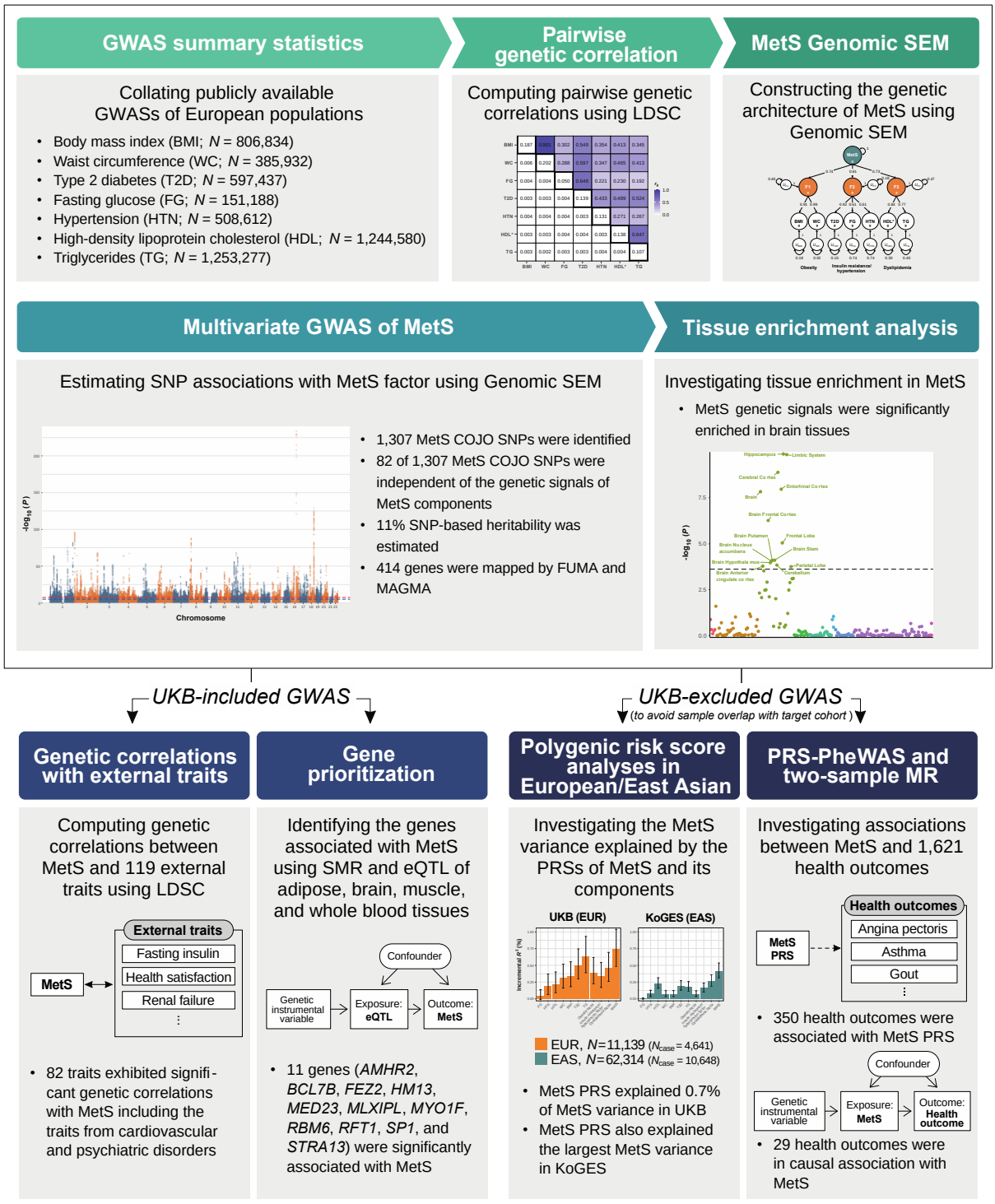
The cML-MA[54] method employs maximum likelihood and model averaging to select valid instrumental variables (IVs), addressing both correlated and uncorrelated pleiotropic effects. Unlike MR-Egger, it does not assume Instrument Strength Independent of Direct Effect (InSIDE) and shows improved control over type I errors. Utilizing the Bayesian information criterion (BIC), it assigns weights and performs model averaging to accommodate model selection uncertainties. Data perturbation can supplement this approach by reducing the risk of overlooking weakly pleiotropic IVs. While cML-MA-BIC-DP is more conservative and offers better type-I error control, it requires more computational resources and may entail decreased power compared with cML-MA-BIC. The choice between cML-MA-BIC and cML-MA-BIC-DP relied on two goodness-of-fit (GOF) tests to assess whether the variance estimates of both methods were equivalent. cML-MA-BIC-DP is preferred if the null hypothesis is rejected; cML-MA-BIC-DP is preferred. The parameters used to perform the cML-MA were *random_start* = 20, *random_start_pert* = 20, *random_seed* = 12345, and *num_pert* = 200.

The cML-MA method was used to assess the robustness of 29 TSMR outcomes, as outlined in **Table 3**. The identified counts of invalid IVs ranged from zero to seven. Across all

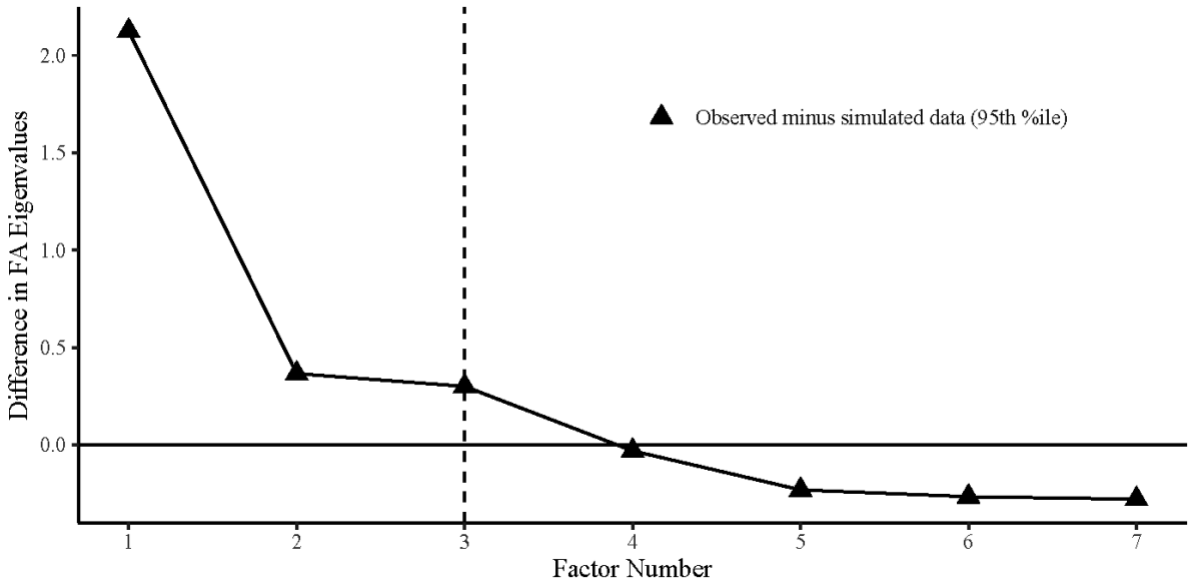
TSMR outcomes represented in **Table 3**, a notable causal relationship was observed using cML-MA, with consistent odds ratio (OR) estimates, except for peripheral angiopathy in diseases classified elsewhere (Phecode = 443.7) (**Supplementary Fig. 13–14**). This deviation might be attributed to the relatively small number of cases (n_{case} = 385) compared to other outcomes, ranging from 1,333 to 38,715.

Supplementary Figures

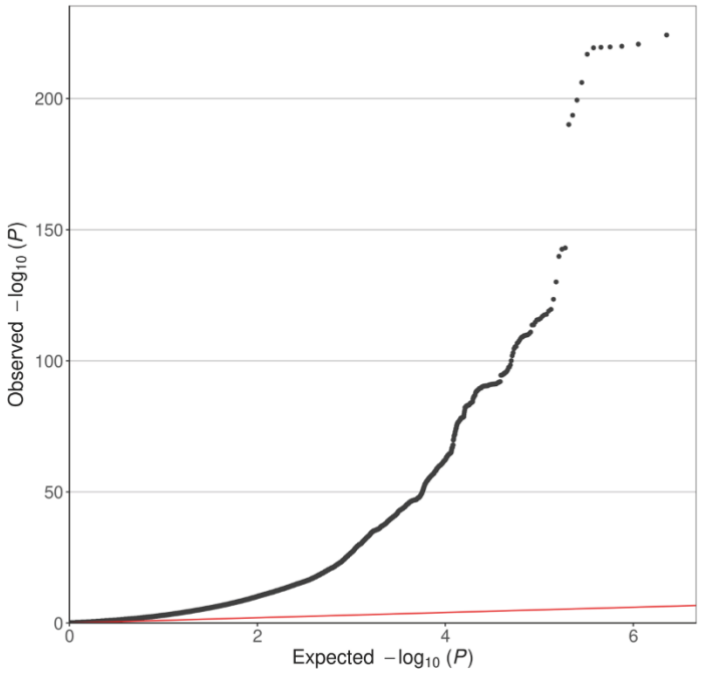
Supplementary Figure 1. Overview of multivariate GWAS of MetS
The workflow of this study is illustrated.



Supplementary Figure 2. Scree plot from parallel analysis
The scree plot was obtained by leveraging the genetic correlation matrix of seven MetS components from LD score regression. The *x*-axis represents the number of factors, and the *y*-axis represents the difference between the eigenvalues computed from the LDSC-derived genetic correlation matrix and the Monte-Carlo simulated genetic correlation matrix. The dotted black line indicates the suggested number of factors to retain for exploratory factor analysis.

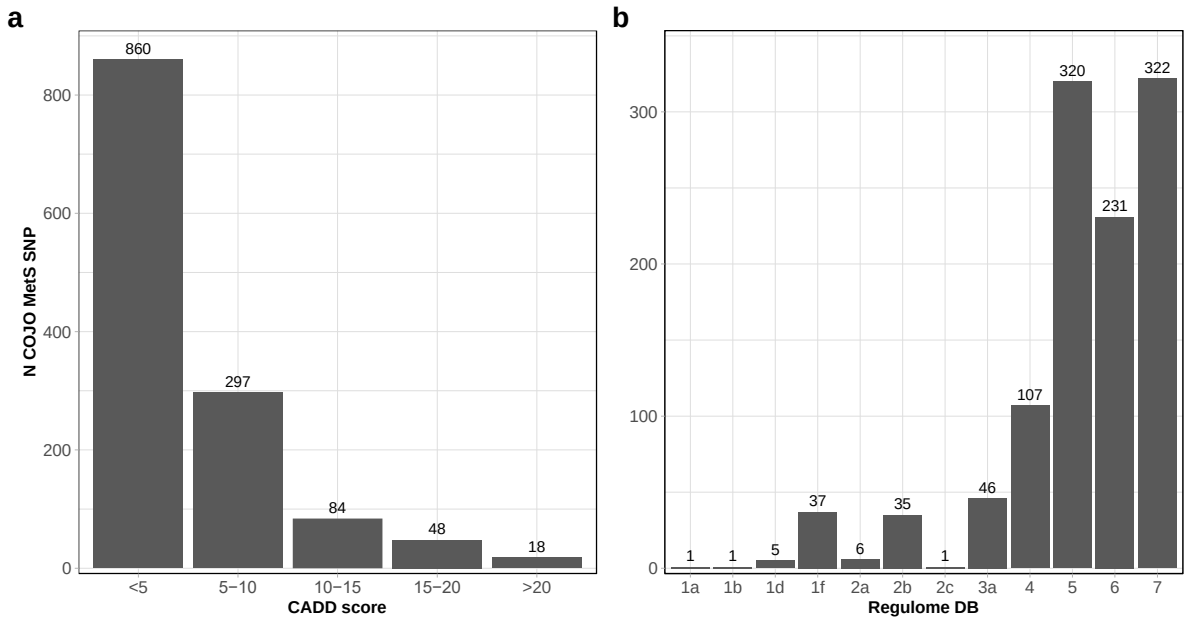


Supplementary Figure 3. Quantile-Quantile (QQ) plot for multivariate MetS GWAS
The *x*- and *y*-axis represent the expected and observed uncorrected two-sided $-\log_{10}(P)$, respectively, for the associations between SNPs and MetS.

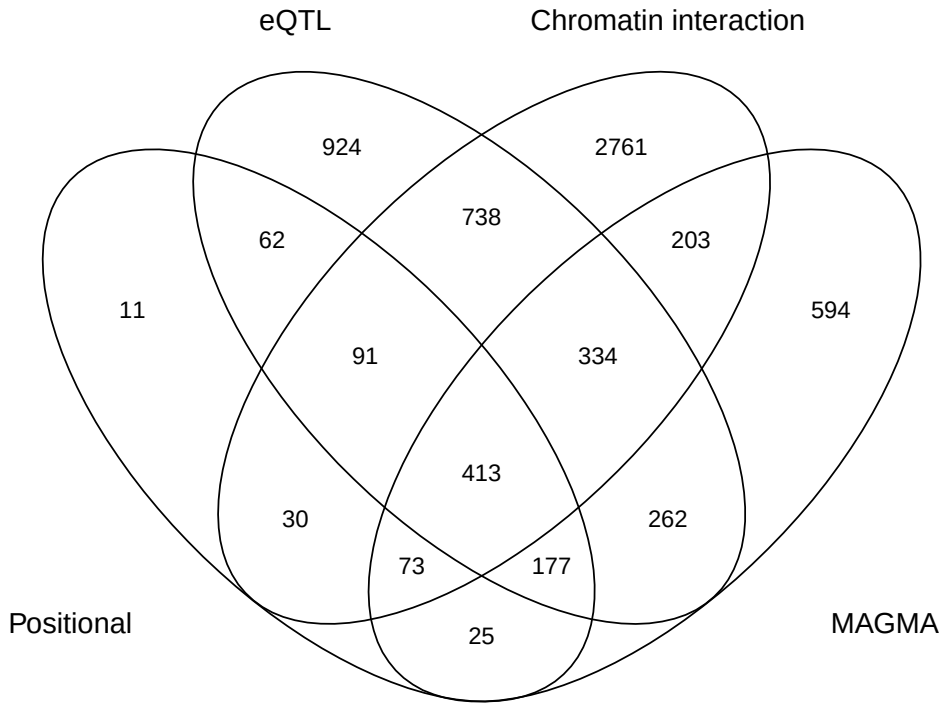


Supplementary Figure 4. Distribution of COJO MetS SNPs annotated with CADD scores and RegulomeDB categories

a, Distribution of CADD scores annotated for 1,307 COJO MetS SNPs. **b**, Distribution of 1,307 COJO MetS SNPs annotated with RegulomeDB categories.

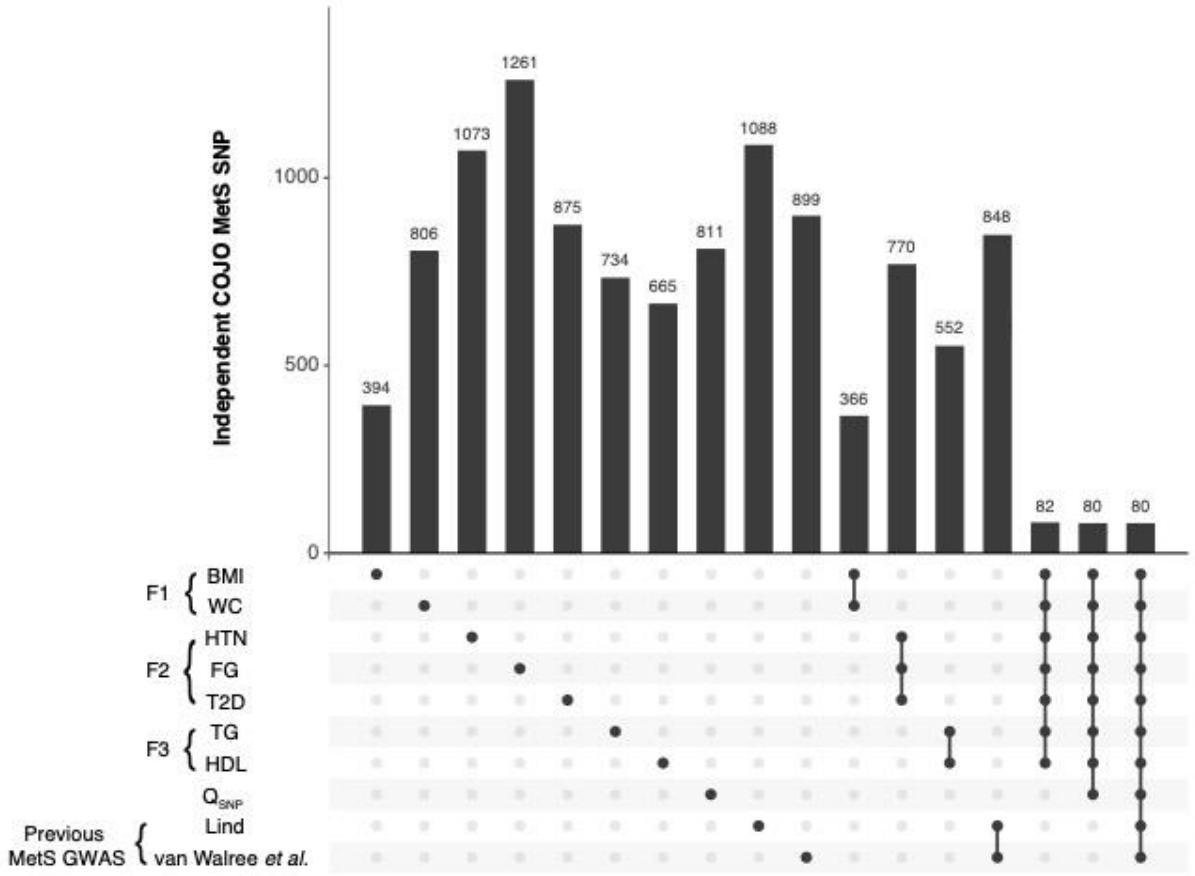


Supplementary Figure 5. **Venn diagram for MetS SNP to gene mapping analysis**
A Venn diagram showing the overlap of genes mapped from MetS GWAS using three different strategies (i.e., positional mapping, eQTL mapping, and chromatin interaction mapping) from FUMA and MAGMA gene-based analysis.

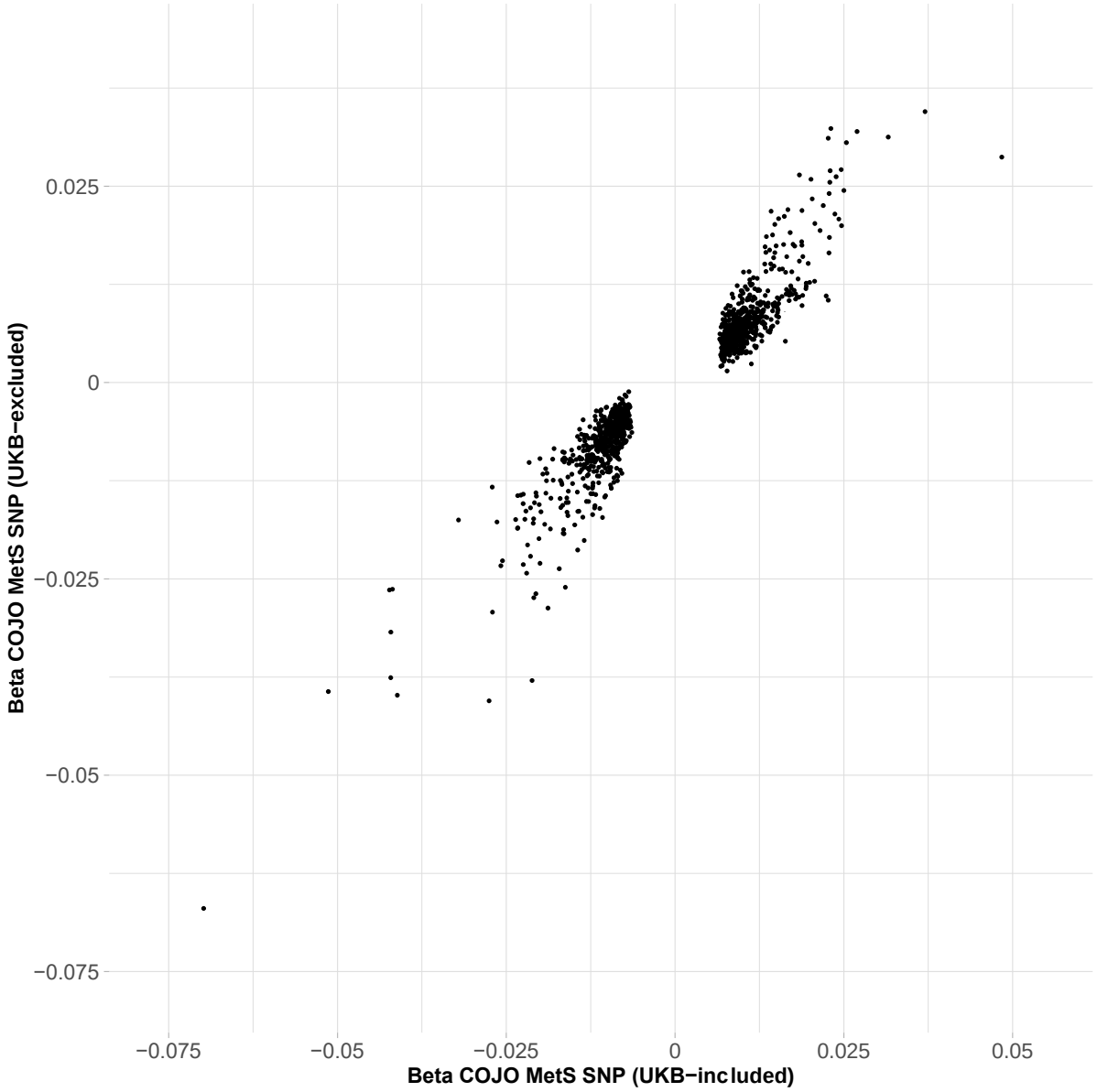


Supplementary Figure 6. COJO MetS SNPs independent from MetS components, Q_{SNP}, and previous MetS studies

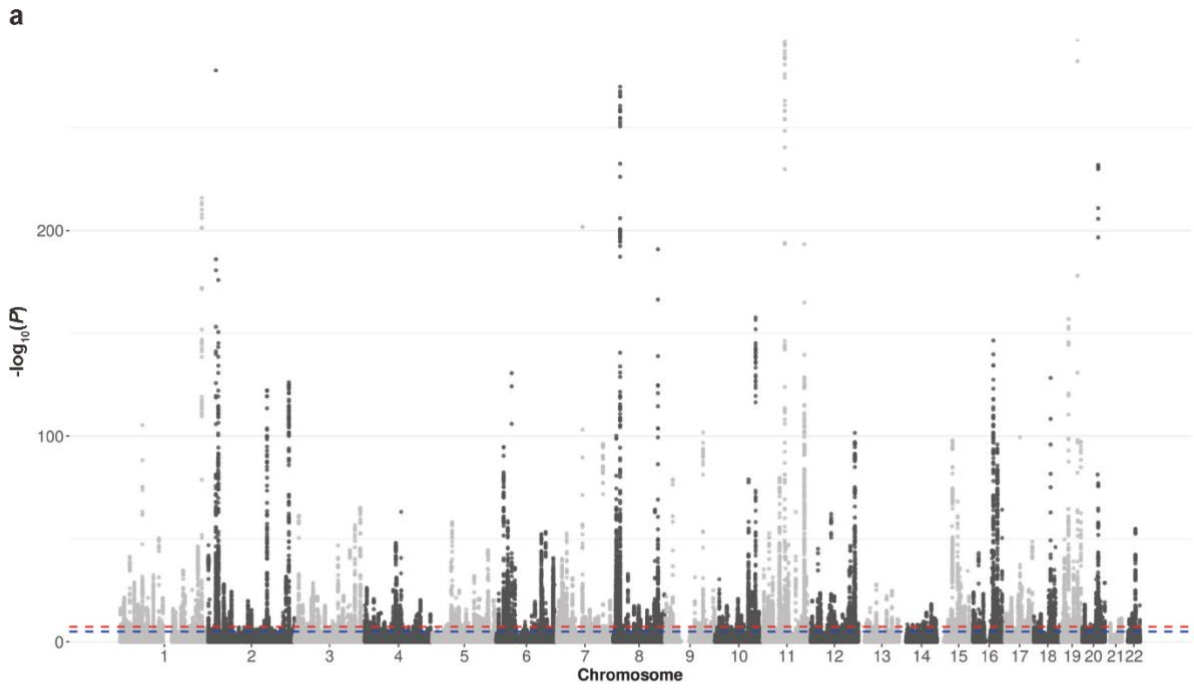
The UpSet plot presents the number of COJO MetS SNPs that were independent of the lead SNPs of GWAS from MetS components, Q_{SNP}, and previous MetS studies. Bar chart shows the number of independent COJO MetS SNPs for each trait. The connected dots in the bottom panel represent traits that were considered simultaneously to determine the independence of the COJO MetS SNPs.

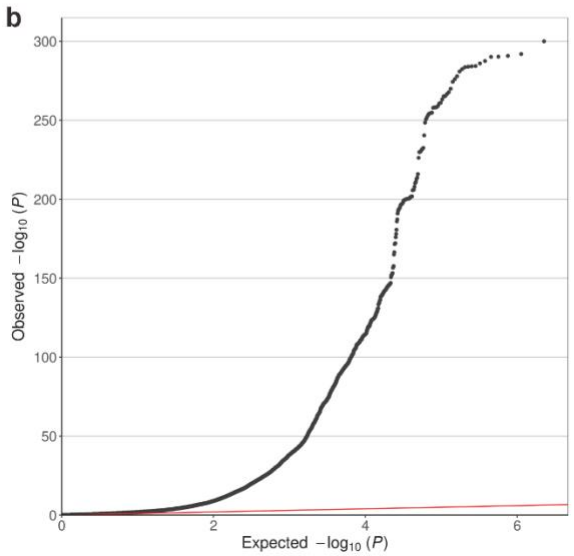


Supplementary Figure 7. GWAS effect estimates concordance of COJO MetS SNPs between UKB-included and UKB-excluded cohorts.
The scatter plot of GWAS effect estimates from MetS GWAS of UKB-included cohorts and UKB-excluded cohorts for 1,307 COJO MetS SNPs identified from GWAS of UKB-included cohorts. The Pearson's correlation between the GWAS effects is 0.95 (*P* from two-sided *z*-test $<2.2 \times 10^{-16}$). The *x*-axis represents GWAS effect estimates from UKB-included cohorts and the *y*-axis represents GWAS effect estimates from UKB-excluded cohorts. The red dashed line is the identity line.



Supplementary Figure 8. Manhattan and Quantile-Quantile (QQ) plot for heterogeneity test (i.e., Q_{SNP}) of MetS GWAS.
a, The Manhattan plot of Q_{SNP} from heterogeneity test for the MetS factor from genomic SEM. The *x*-axis represents chromosomal position, and the *y*-axis represents uncorrected $-\log_{10}(P)$ from one-sided χ^2-test with 2 degrees of freedom for Q_{SNP} test of heterogeneity. **b**, The *x*- and *y*-axis represent the expected and observed uncorrected $-\log_{10}(P)$, respectively, for Q_{SNP} test of heterogeneity for MetS.

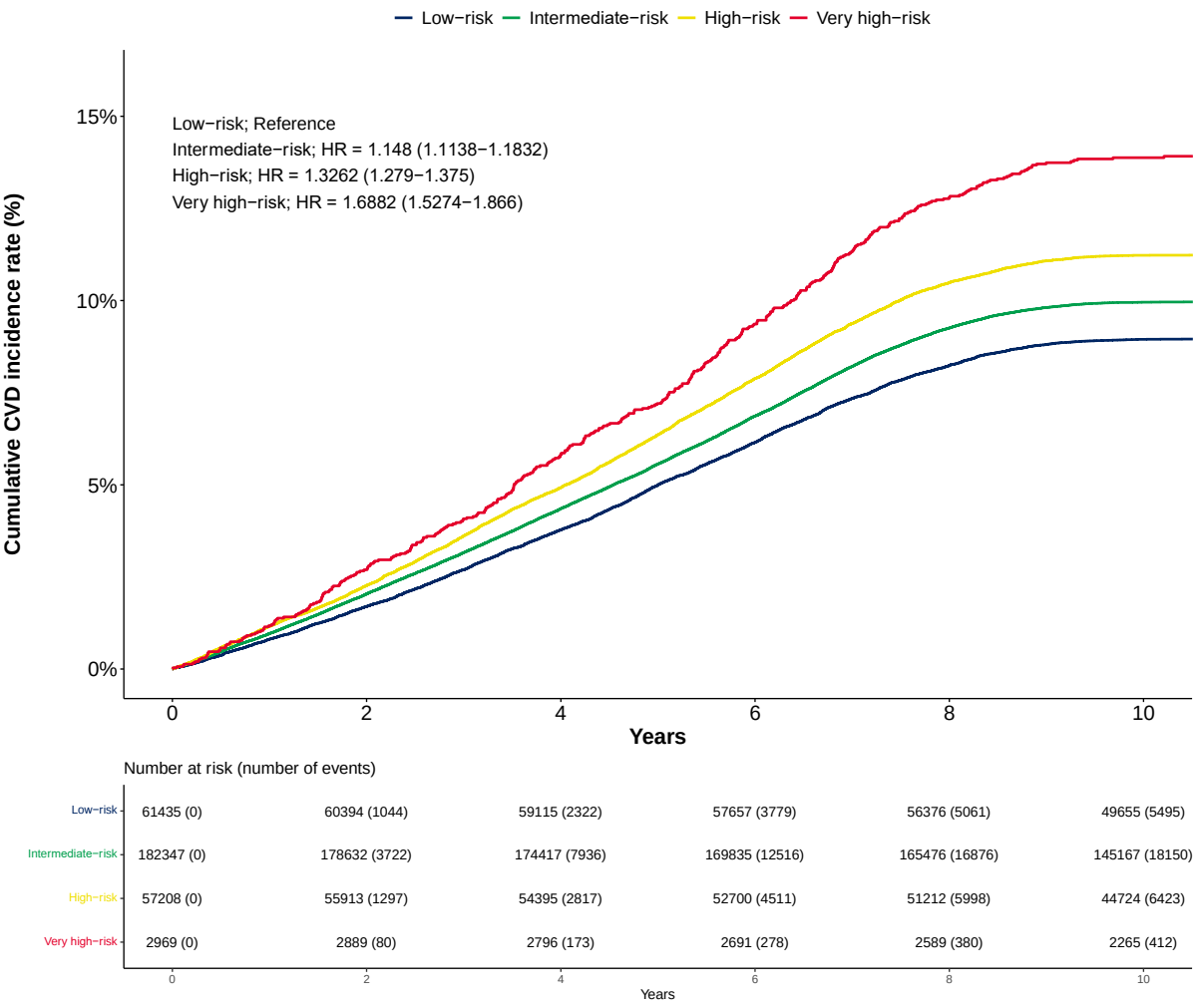




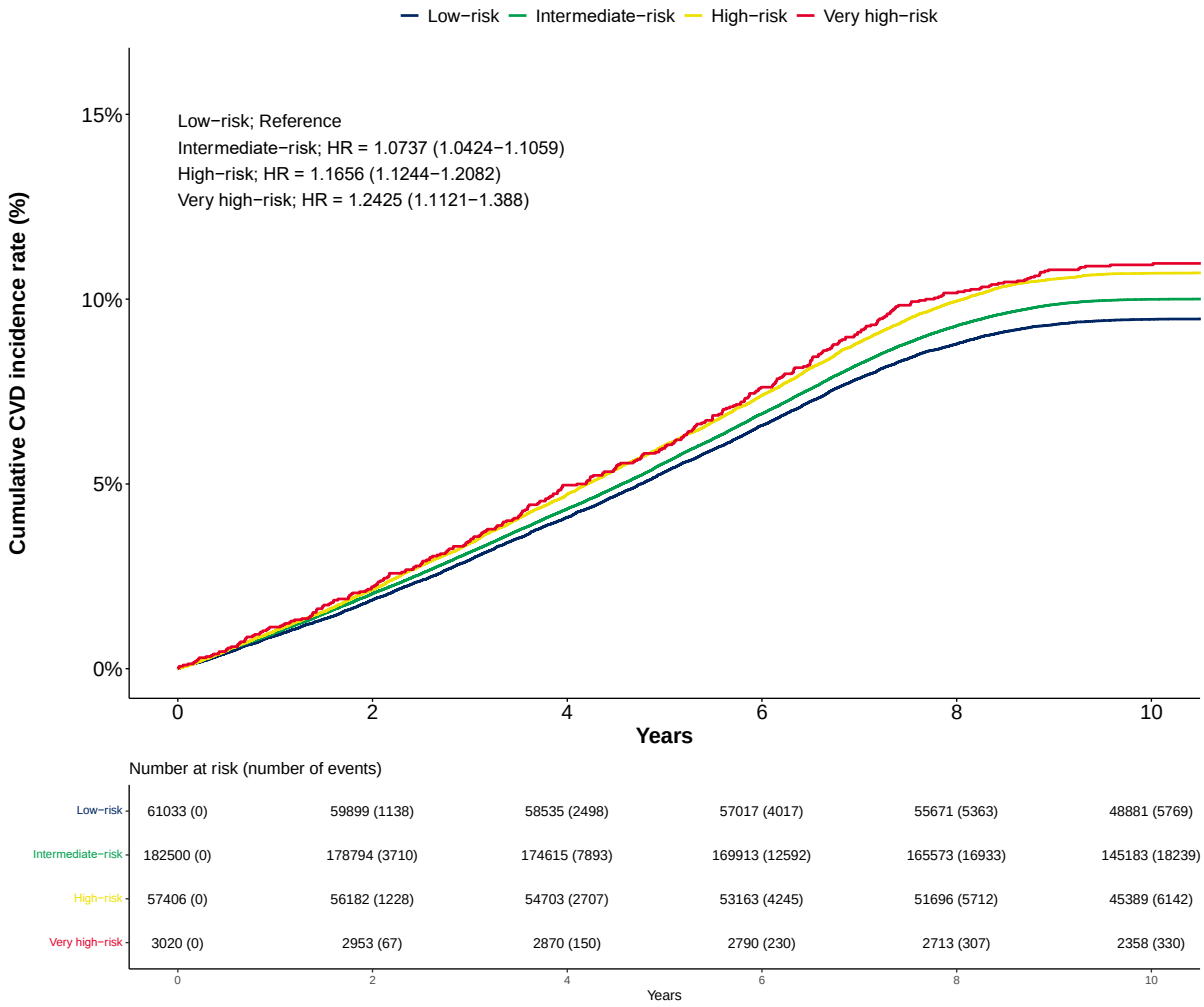
Supplementary Figure 9. Hazard plot for PRS of MetS and its components with cardiovascular disease incidence rate in the UK Biobank

Hazard plot for cardiovascular disease incidence rate in UK Biobank (UKB) with PRS stratified into four groups: low risk, intermediate risk, high risk, and very high risk. The HR of the intermediate-risk, high-risk, and very high-risk groups were annotated with low-risk as the reference group. **a**, MetS PRS. **b**, BMI PRS. **c**, WC PRS. **d**, HTN PRS. **e**, FG PRS. **f**, T2D PRS. **g**, HDL* PRS. **h**, TG PRS. HR, hazard ratio; MetS, metabolic syndrome; BMI, body mass index; WC, waist circumference; FG, fasting glucose; HTN, hypertension; T2D, type 2 diabetes; HDL*: high-density lipoprotein; TG, triglyceride. *Reverse-coded.

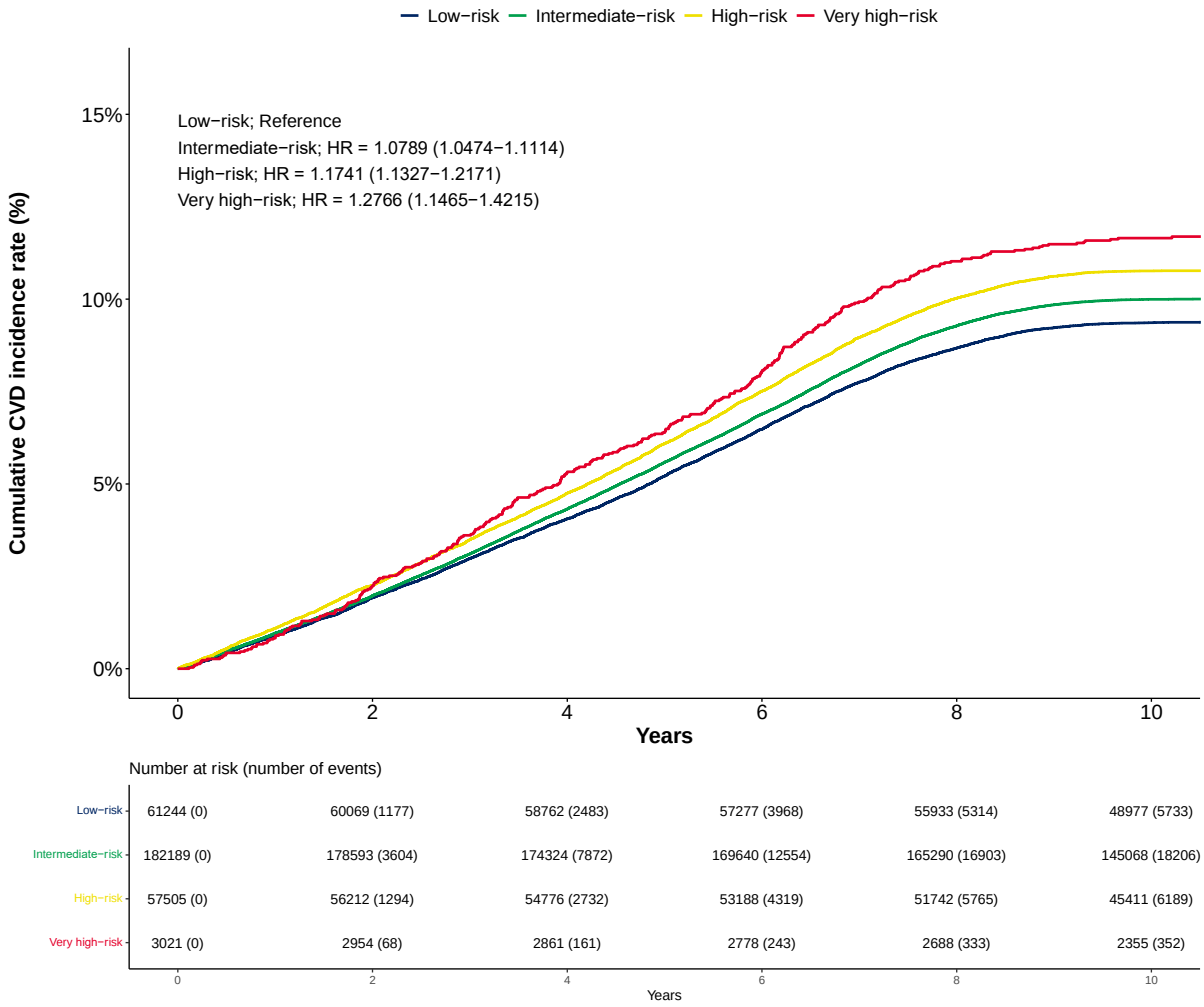
a. MetS PRS



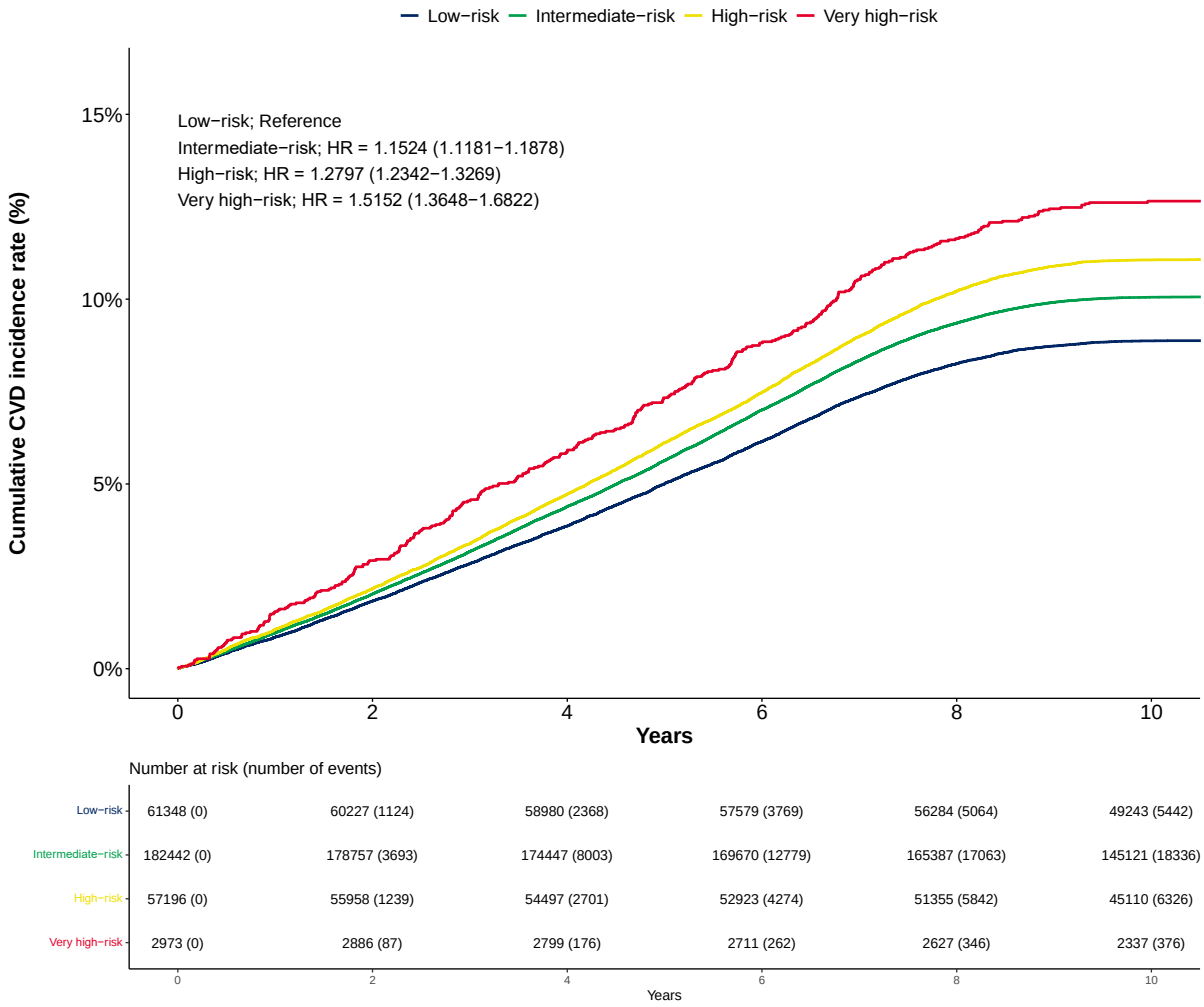
b. BMI PRS


Low−risk
Intermediate−risk
High−risk
Very high−risk
Low−risk; Reference
Intermediate−risk; HR = 1.0737 (1.0424−1.1059)
High−risk; HR = 1.1656 (1.1244−1.2082)
Very high−risk; HR = 1.2425 (1.1121−1.388)
Cumulative CVD incidence rate (%)
15%
10%
5%
0%
0
2
4
6
8
10
Years
Number at risk (number of events)
Low−risk 61033 (0) 59899 (1138) 58535 (2498) 57017 (4017) 55671 (5363) 48881 (5769)
Intermediate−risk 182500 (0) 178794 (3710) 174615 (7893) 169913 (12592) 165573 (16933) 145183 (18239)
High−risk 57406 (0) 56182 (1228) 54703 (2707) 53163 (4245) 51696 (5712) 45389 (6142)
Very high−risk 3020 (0) 2953 (67) 2870 (150) 2790 (230) 2713 (307) 2358 (330)
0
2
4
6
8
10
Years

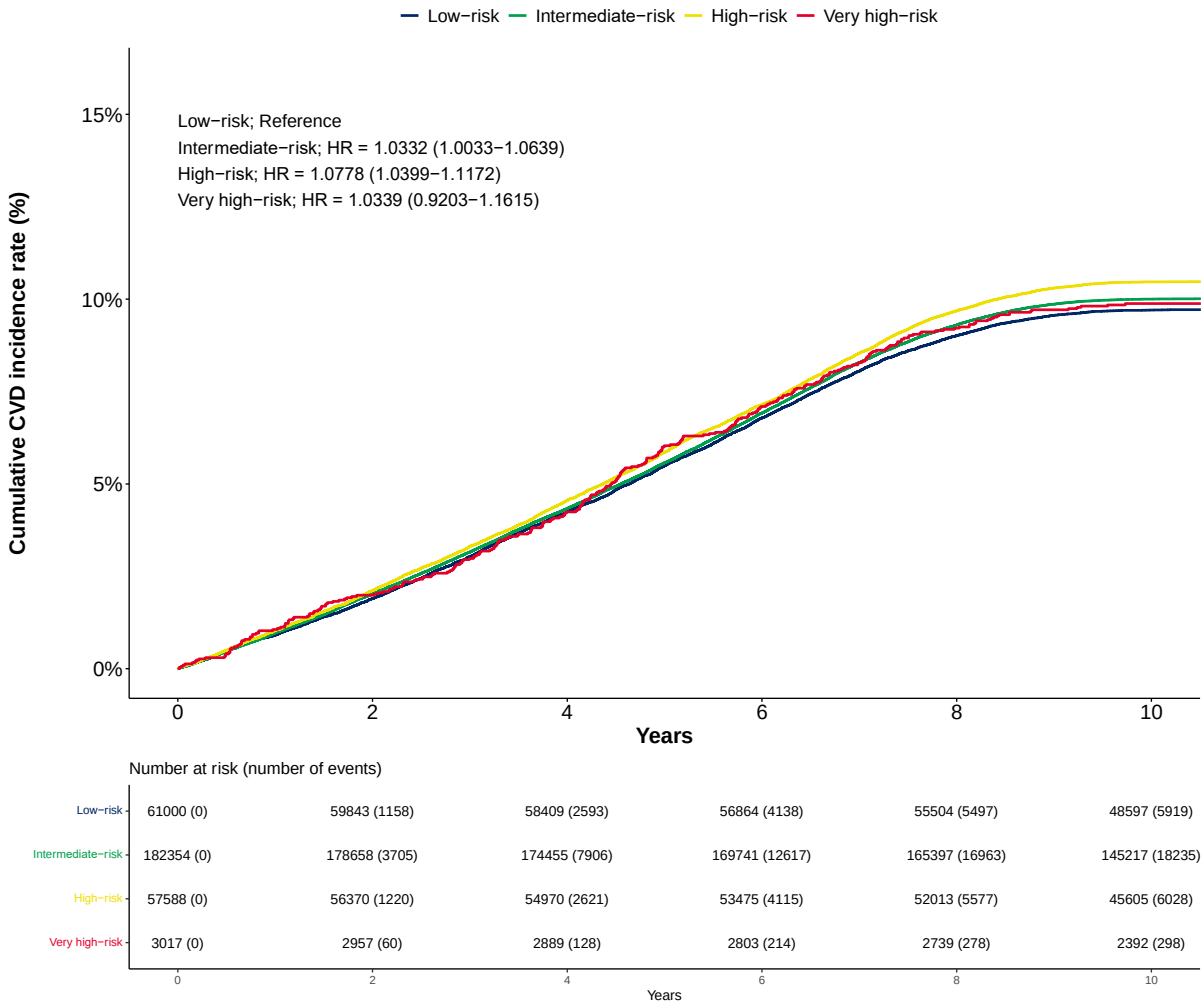
c. WC PRS


Low−risk
Intermediate−risk
High−risk
Very high−risk
Low−risk; Reference
Intermediate−risk; HR = 1.0789 (1.0474−1.1114)
High−risk; HR = 1.1741 (1.1327−1.2171)
Very high−risk; HR = 1.2766 (1.1465−1.4215)
Cumulative CVD incidence rate (%)
15%
10%
5%
0%
0
2
4
6
8
10
Years
Number at risk (number of events)
Low−risk 61244 (0) 60069 (1177) 58762 (2483) 57277 (3968) 55933 (5314) 48977 (5733)
Intermediate−risk 182189 (0) 178593 (3604) 174324 (7872) 169640 (12554) 165290 (16903) 145068 (18206)
High−risk 57505 (0) 56212 (1294) 54776 (2732) 53188 (4319) 51742 (5765) 45411 (6189)
Very high−risk 3021 (0) 2954 (68) 2861 (161) 2778 (243) 2688 (333) 2355 (352)
0
2
4
6
8
10
Years

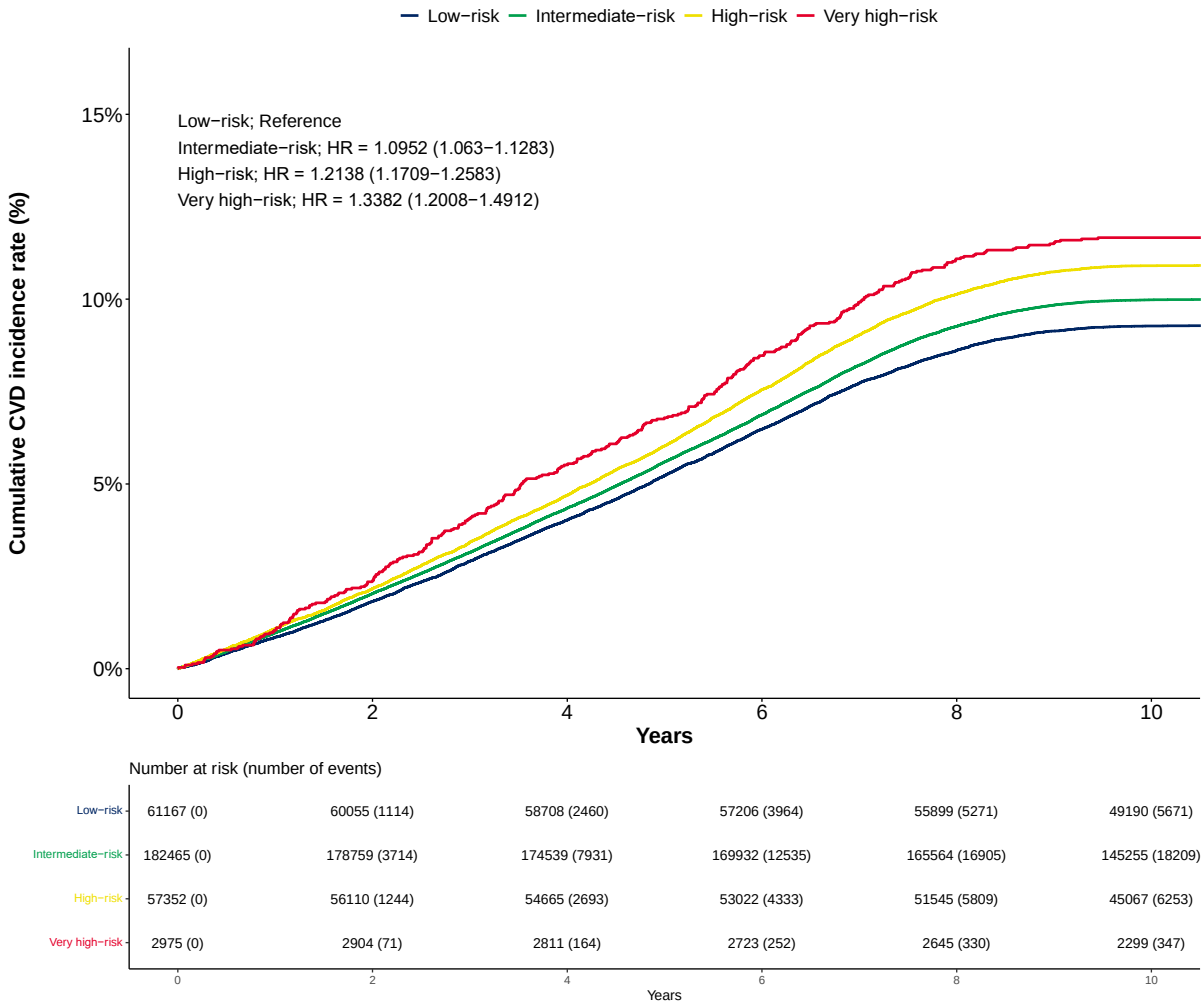
d. HTN PRS


Low−risk
Intermediate−risk
High−risk
Very high−risk
Low−risk; Reference
Intermediate−risk; HR = 1.1524 (1.1181−1.1878)
High−risk; HR = 1.2797 (1.2342−1.3269)
Very high−risk; HR = 1.5152 (1.3648−1.6822)
Cumulative CVD incidence rate (%)
15%
10%
5%
0%
0
2
4
6
8
10
Years
Number at risk (number of events)
Low−risk 61348 (0) 60227 (1124) 58980 (2368) 57579 (3769) 56284 (5064) 49243 (5442)
Intermediate−risk 182442 (0) 178757 (3693) 174447 (8003) 169670 (12779) 165387 (17063) 145121 (18336)
High−risk 57196 (0) 55958 (1239) 54497 (2701) 52923 (4274) 51355 (5842) 45110 (6326)
Very high−risk 2973 (0) 2886 (87) 2799 (176) 2711 (262) 2627 (346) 2337 (376)
0
2
4
6
8
10
Years

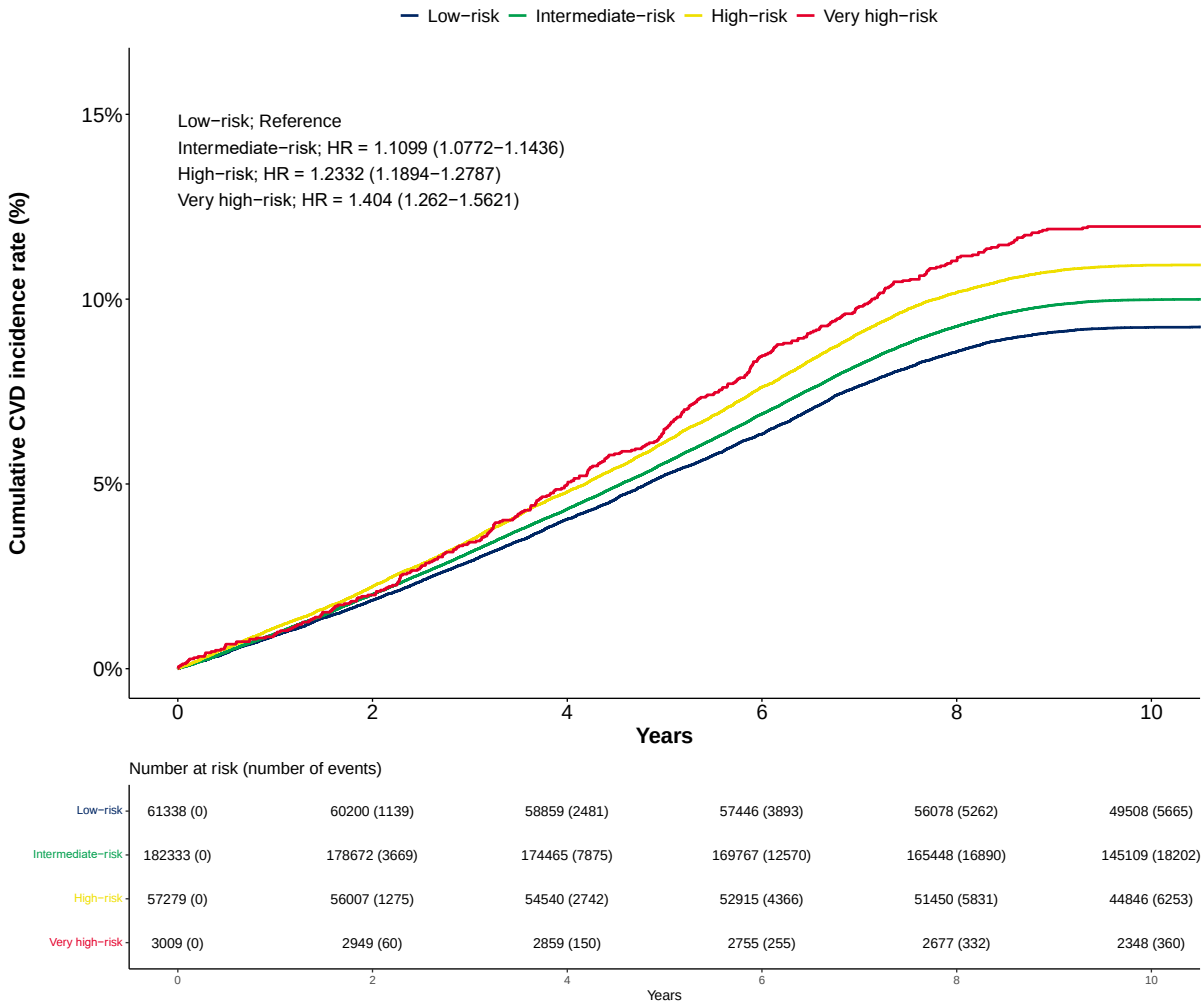
e. FG PRS


Low−risk
Intermediate−risk
High−risk
Very high−risk
Low−risk; Reference
Intermediate−risk; HR = 1.0332 (1.0033−1.0639)
High−risk; HR = 1.0778 (1.0399−1.1172)
Very high−risk; HR = 1.0339 (0.9203−1.1615)
Cumulative CVD incidence rate (%)
15%
10%
5%
0%
0
2
4
6
8
10
Years
Number at risk (number of events)
Low−risk 61000 (0) 59843 (1158) 58409 (2593) 56864 (4138) 55504 (5497) 48597 (5919)
Intermediate−risk 182354 (0) 178658 (3705) 174455 (7906) 169741 (12617) 165397 (16963) 145217 (18235)
High−risk 57588 (0) 56370 (1220) 54970 (2621) 53475 (4115) 52013 (5577) 45605 (6028)
Very high−risk 3017 (0) 2957 (60) 2889 (128) 2803 (214) 2739 (278) 2392 (298)
0
2
4
6
8
10
Years

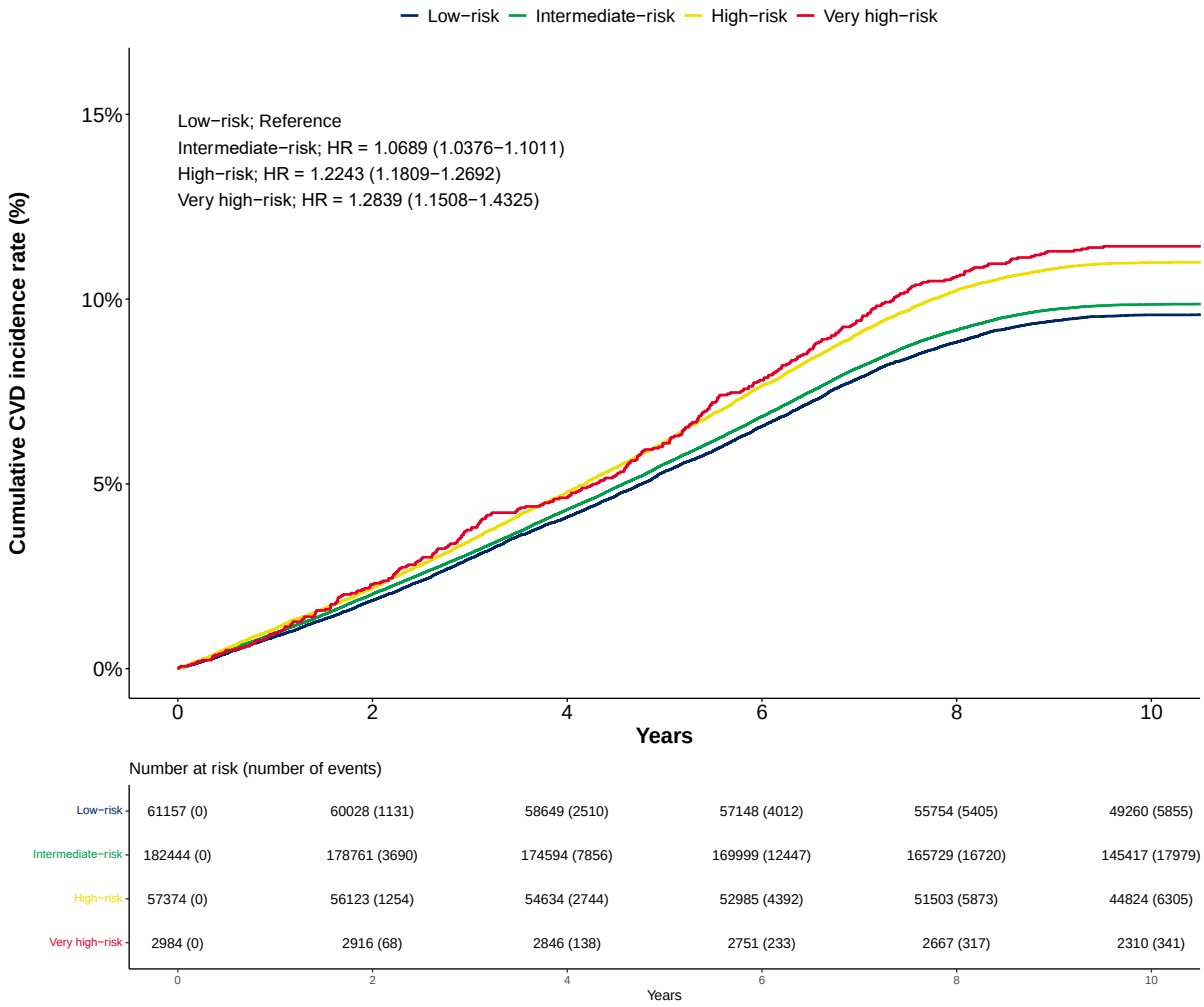
f. T2D PRS


Low−risk
Intermediate−risk
High−risk
Very high−risk
Low−risk; Reference
Intermediate−risk; HR = 1.0952 (1.063−1.1283)
High−risk; HR = 1.2138 (1.1709−1.2583)
Very high−risk; HR = 1.3382 (1.2008−1.4912)
Cumulative CVD incidence rate (%)
15%
10%
5%
0%
0
2
4
6
8
10
Years
Number at risk (number of events)
Low−risk 61167 (0) 60055 (1114) 58708 (2460) 57206 (3964) 55899 (5271) 49190 (5671)
Intermediate−risk 182465 (0) 178759 (3714) 174539 (7931) 169932 (12535) 165564 (16905) 145255 (18209)
High−risk 57352 (0) 56110 (1244) 54665 (2693) 53022 (4333) 51545 (5809) 45067 (6253)
Very high−risk 2975 (0) 2904 (71) 2811 (164) 2723 (252) 2645 (330) 2299 (347)
0
2
4
6
8
10
Years

g. HDL* PRS

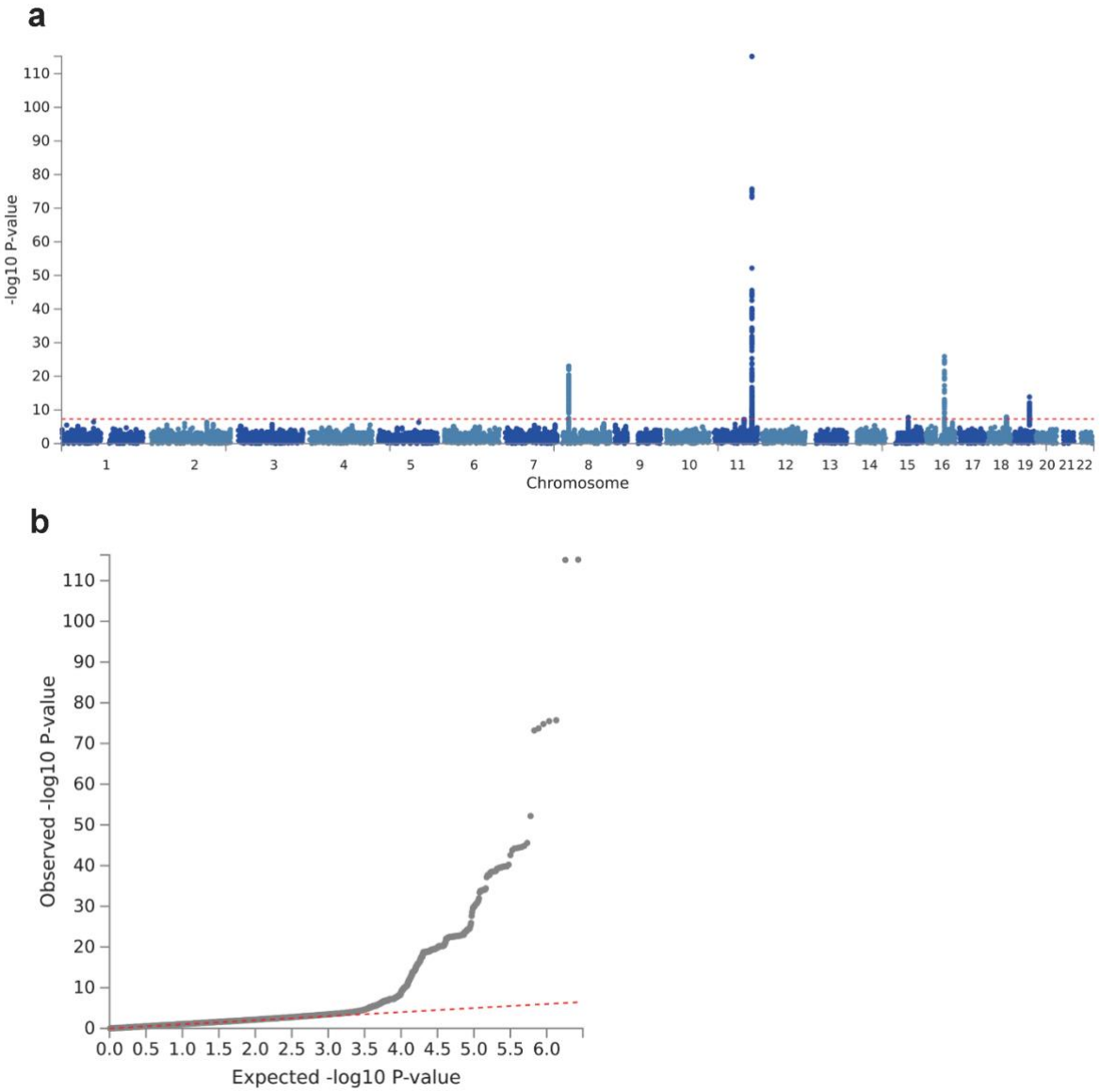

Low−risk
Intermediate−risk
High−risk
Very high−risk
Low−risk; Reference
Intermediate−risk; HR = 1.1099 (1.0772−1.1436)
High−risk; HR = 1.2332 (1.1894−1.2787)
Very high−risk; HR = 1.404 (1.262−1.5621)
Cumulative CVD incidence rate (%)
15%
10%
5%
0%
0
2
4
6
8
10
Years
Number at risk (number of events)
Low−risk 61338 (0) 60200 (1139) 58859 (2481) 57446 (3893) 56078 (5262) 49508 (5665)
Intermediate−risk 182333 (0) 178672 (3669) 174465 (7875) 169767 (12570) 165448 (16890) 145109 (18202)
High−risk 57279 (0) 56007 (1275) 54540 (2742) 52915 (4366) 51450 (5831) 44846 (6253)
Very high−risk 3009 (0) 2949 (60) 2859 (150) 2755 (255) 2677 (332) 2348 (360)
0
2
4
6
8
10
Years

h. TG PRS


Low−risk
Intermediate−risk
High−risk
Very high−risk
Low−risk; Reference
Intermediate−risk; HR = 1.0689 (1.0376−1.1011)
High−risk; HR = 1.2243 (1.1809−1.2692)
Very high−risk; HR = 1.2839 (1.1508−1.4325)
Cumulative CVD incidence rate (%)
15%
10%
5%
0%
0
2
4
6
8
10
Years
Number at risk (number of events)
Low−risk 61157 (0) 60028 (1131) 58649 (2510) 57148 (4012) 55754 (5405) 49260 (5855)
Intermediate−risk 182444 (0) 178761 (3690) 174594 (7856) 169999 (12447) 165729 (16720) 145417 (17979)
High−risk 57374 (0) 56123 (1254) 54634 (2744) 52985 (4392) 51503 (5873) 44824 (6305)
Very high−risk 2984 (0) 2916 (68) 2846 (138) 2751 (233) 2667 (317) 2310 (341)
0
2
4
6
8
10
Years

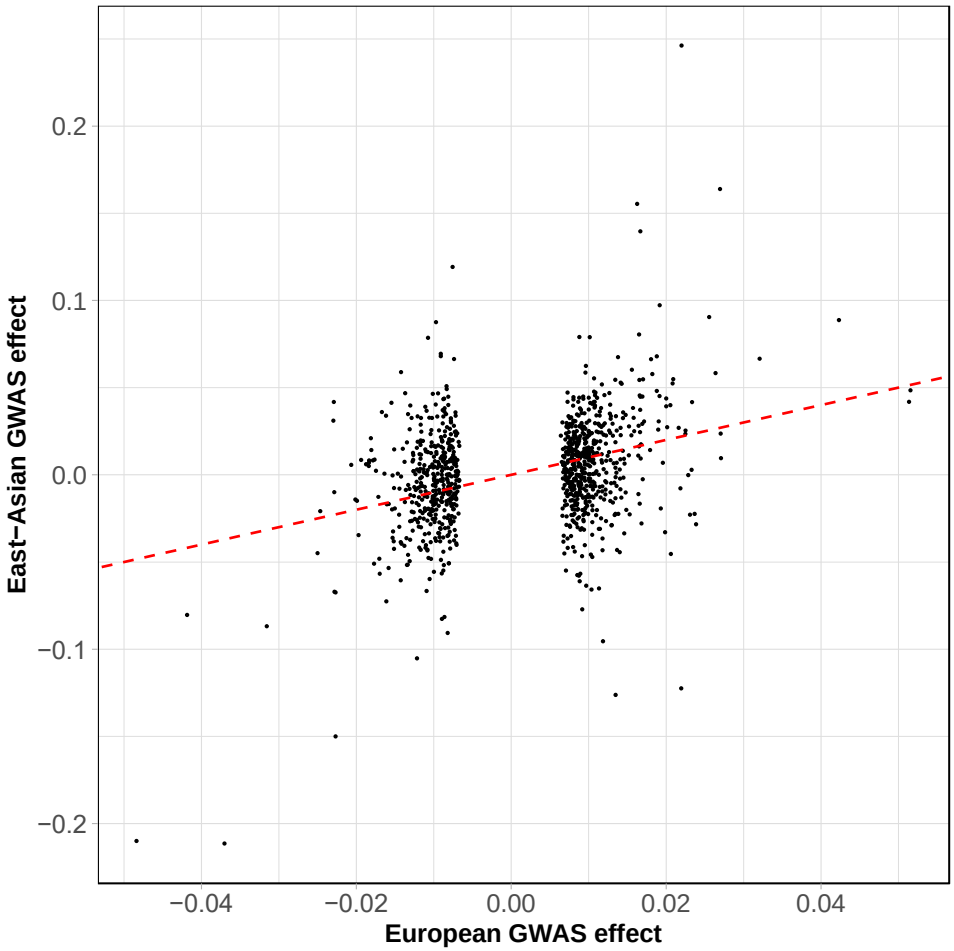
Supplementary Figure 10. Manhattan and Quantile-Quantile (QQ) plot for MetS GWAS in East Asian population

a, The Manhattan plot of dichotomized MetS GWAS in East-Asian population. The *x*-axis represents chromosomal position, and the *y*-axis represents uncorrected $-\log_{10}(P)$ for two-sided *z*-test for the SNP association with MetS. **b**, The *x*- and *y*-axis represent the expected and observed uncorrected two-sided $-\log_{10}(P)$, respectively, for the SNP associations with MetS.



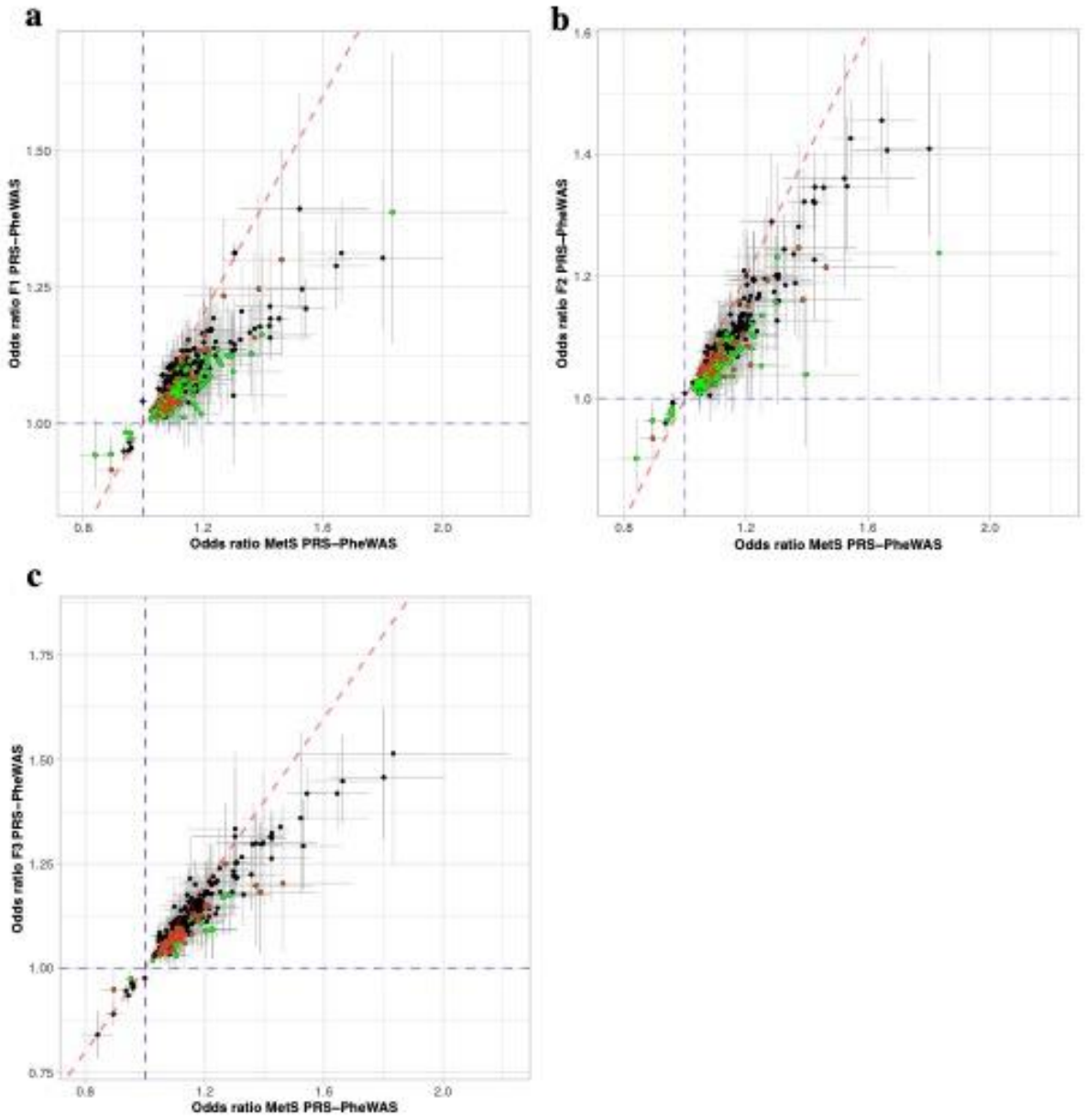
Supplementary Figure 11. GWAS effect estimates concordance of COJO MetS SNPs from the European MetS GWAS to the East Asian MetS GWAS

The scatter plot of GWAS effect estimates from MetS GWAS in the European and East Asian population for 1,016 COJO MetS SNPs identified from European GWAS (Note that the remaining 291 COJO MetS SNPs were unavailable in the East-Asian GWAS). The Pearson's correlation between the GWAS effects is 0.34 (*P* from two-sided *z*-test = 3.85×10^{-28}). The *x*-axis represents MetS GWAS effect estimates from the European population and the *y*-axis represents MetS GWAS effect estimates from the East-Asian population. The red dashed line is the identity line.



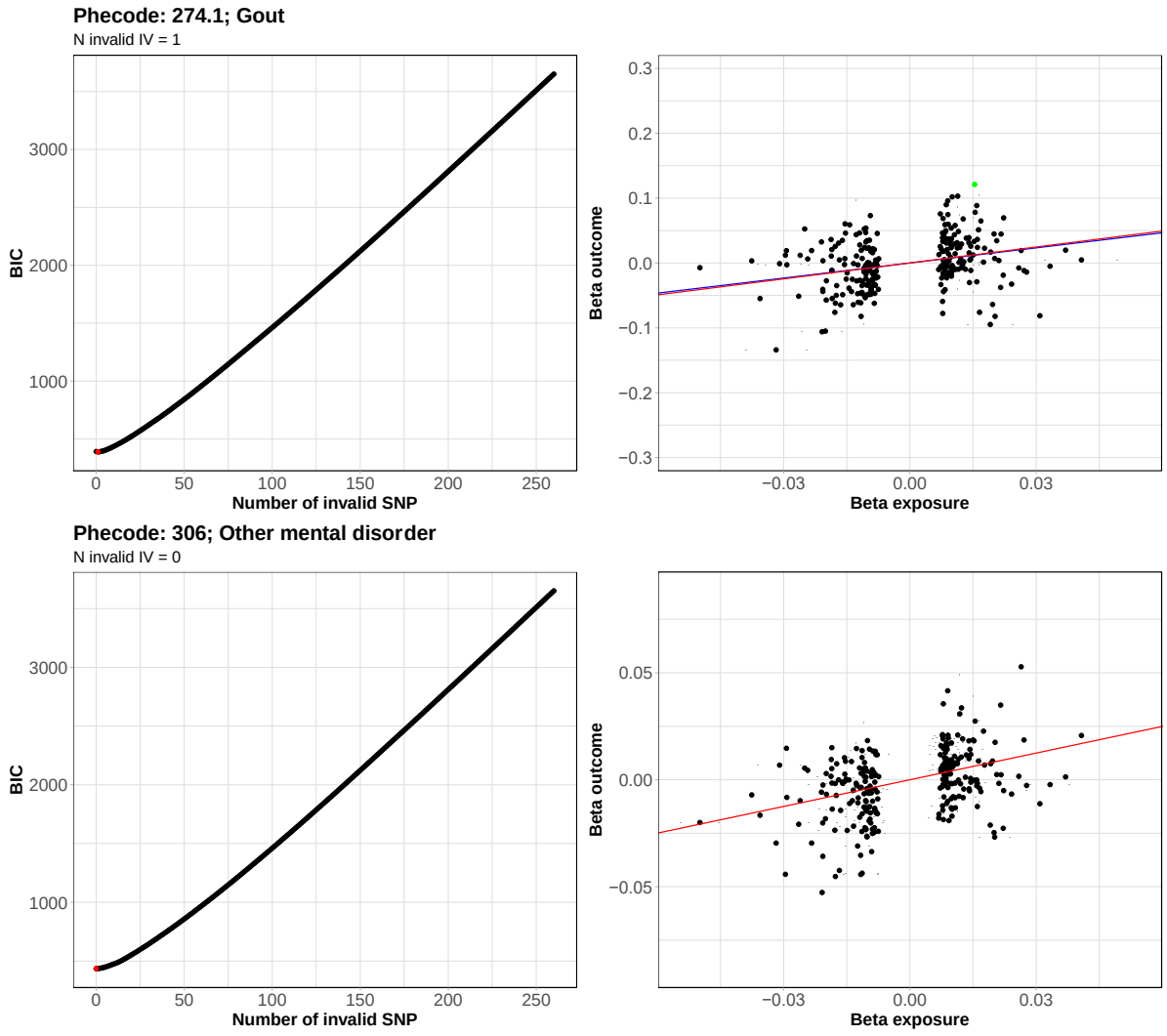
Supplementary Figure 12. Scatter plot of odds ratio (OR) of PRS-PheWAS results that showed significant association with MetS PRS

a, Comparison of OR from PRS-PheWAS between MetS PRS and F1 PRS. **b,** Comparison of OR from PRS-PheWAS between MetS PRS and F2 PRS. **c,** Comparison of OR from PRS-PheWAS between MetS PRS and F3 PRS. In all panels, green dots represent the outcomes that were significantly associated with MetS and the corresponding factor PRS in comparison; red dots represent the outcomes that were significantly associated with only MetS PRS; grey bars are error bars with 95% confidence intervals, computed as OR ± 1.96 × standard error; the blue dashed line represents OR = 1; and the red dashed line represents the identity line. The sample size for each data point can be found in **Supplementary Tables 40** and **41**.



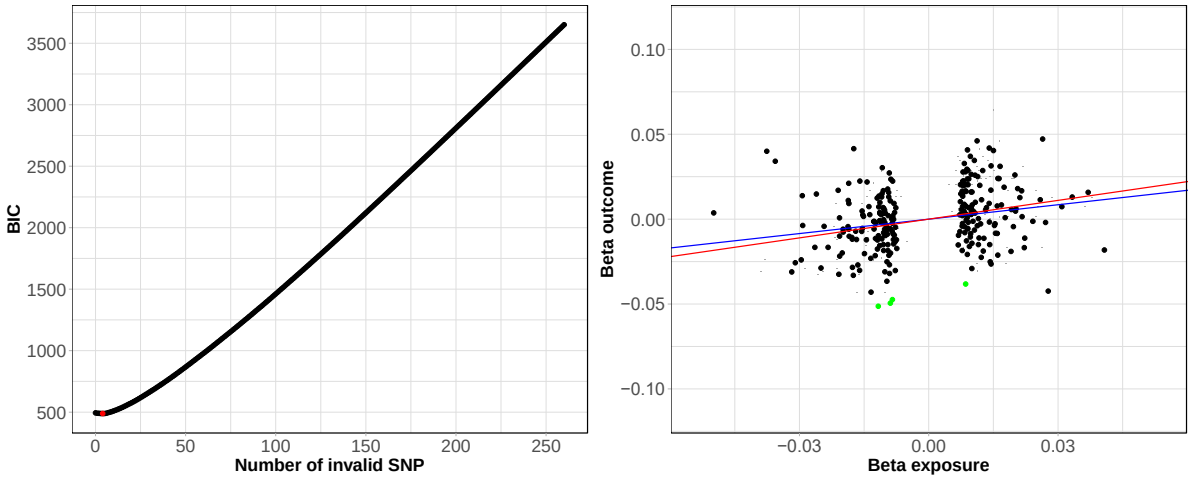
Supplementary Figure 13. BIC and scatter plot pairs for 29 health outcomes were tested for their causal association with MetS using constrained maximum likelihood and model averaging-based Mendelian randomization (cML-MA MR)

The left panel shows the number of invalid instrumental variables (IVs) on the *x*-axis and BIC values on the *y*-axis. The right panel shows the beta from exposure (i.e., MetS) and outcome (i.e., health outcome) on the *x*-axis and *y*-axis, respectively, with their respective error bars. The green dots represent invalid IVs detected using cML-MA, the black dashed lines represent beta = 0, the blue line represents the causal estimate from the inverse variance weighted (IVW) method, and the red line represents the causal estimate from either cML-MA-BIC or cML-MA-BIC-DP MR.



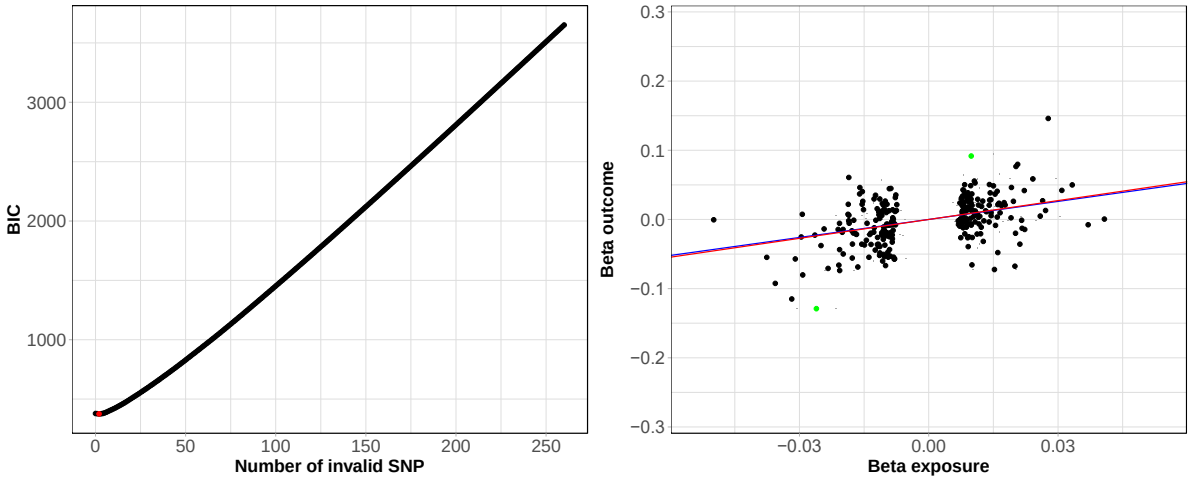
Phecode: 318; Tobacco use disorder

N invalid IV = 4



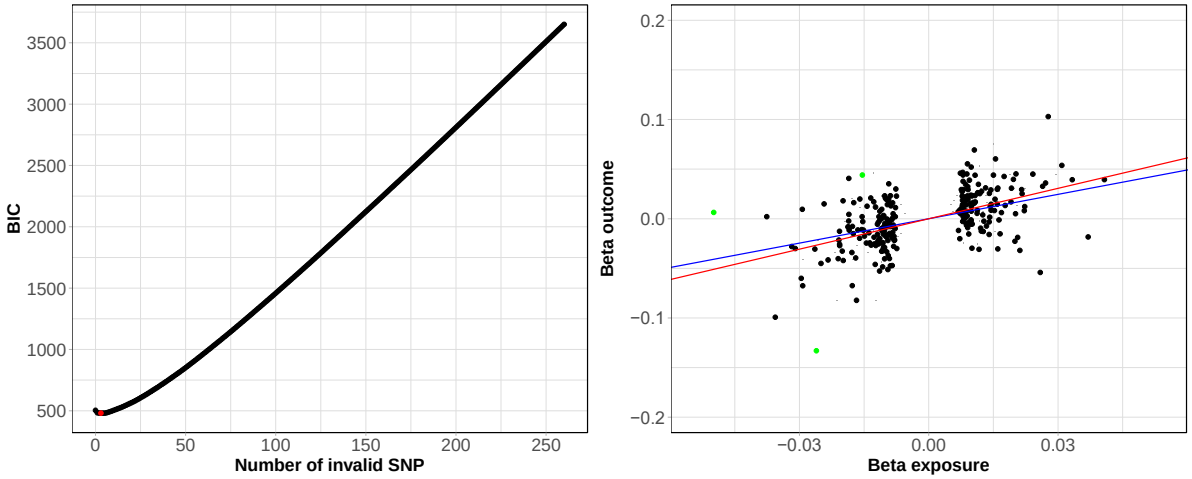
Phecode: 411.1; Unstable angina (intermediate coro

N invalid IV = 2



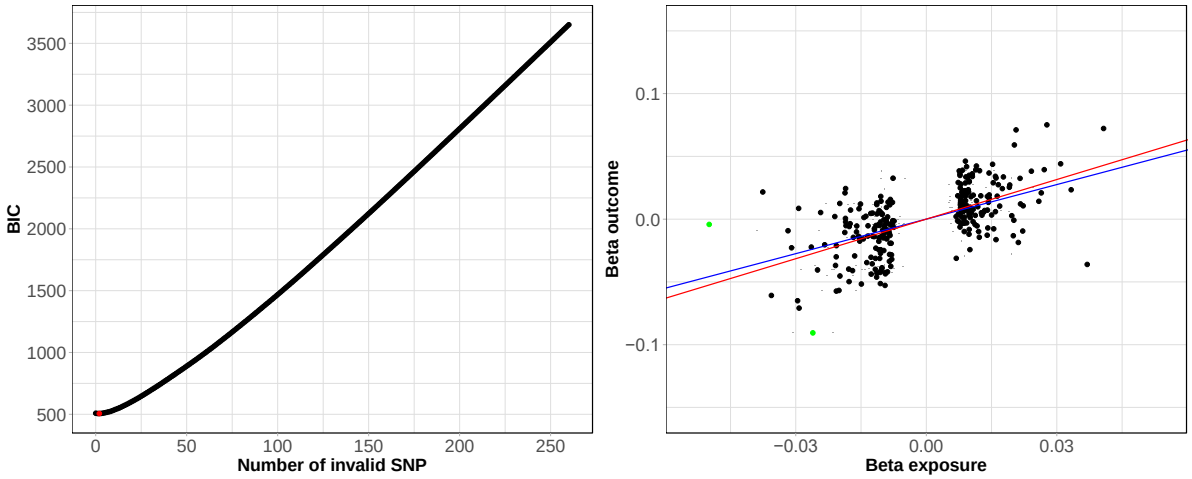
Phecode: 411.2; Myocardial infarction

N invalid IV = 3



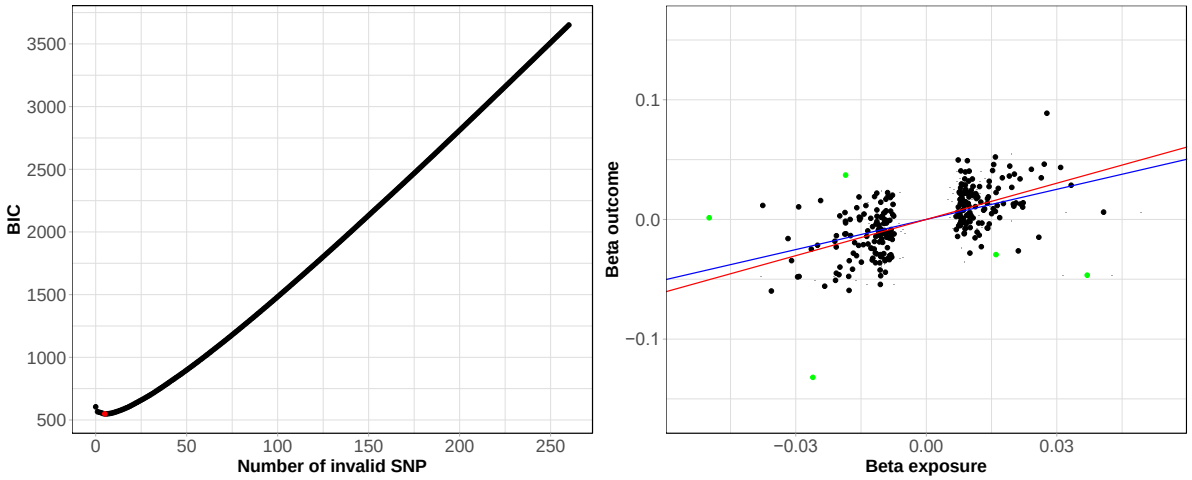
Phecode: 411.3; Angina pectoris

N invalid IV = 2



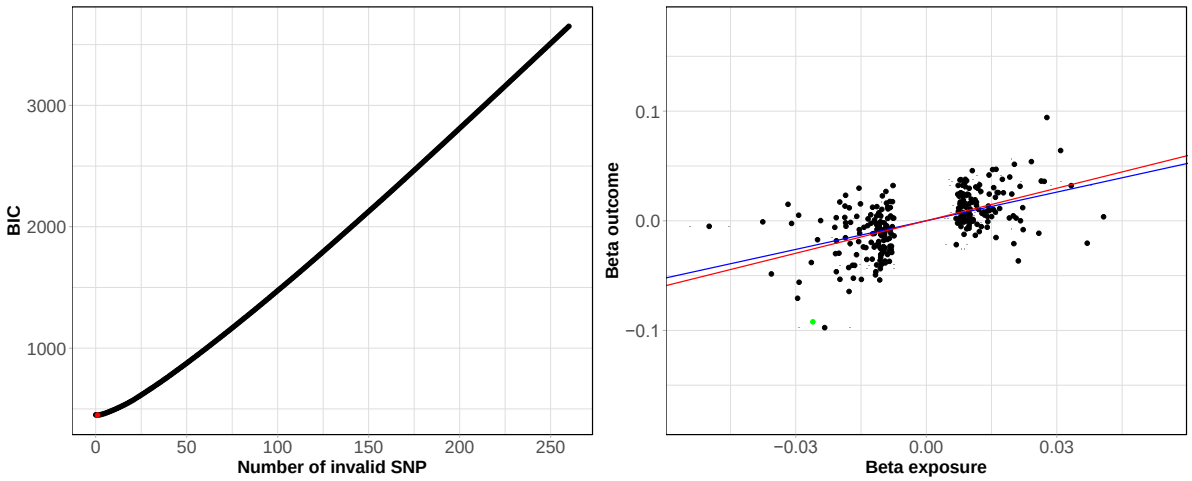
Phecode: 411.4; Coronary atherosclerosis

N invalid IV = 5



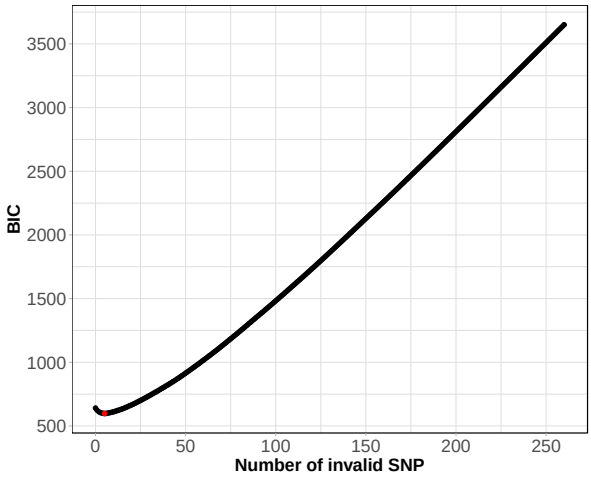
Phecode: 411.8; Other chronic ischemic heart disea

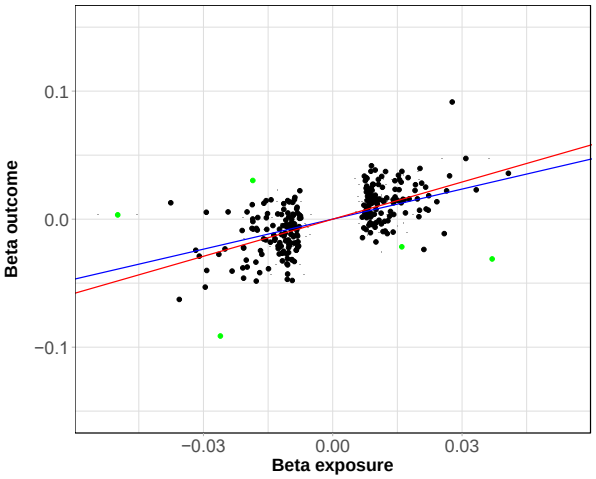
N invalid IV = 1



Phecode: 411; Ischemic Heart Disease

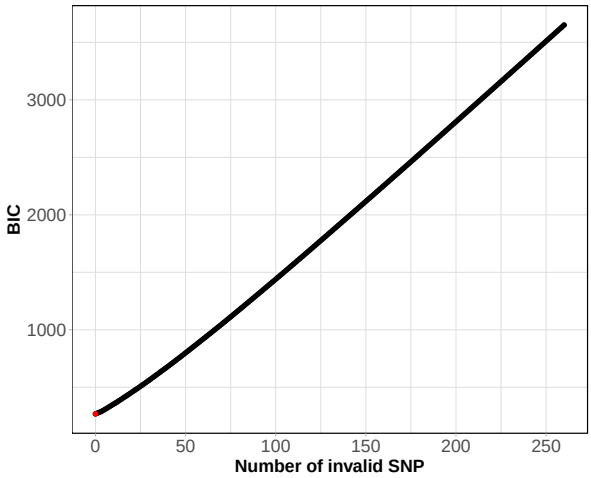
N invalid IV = 5

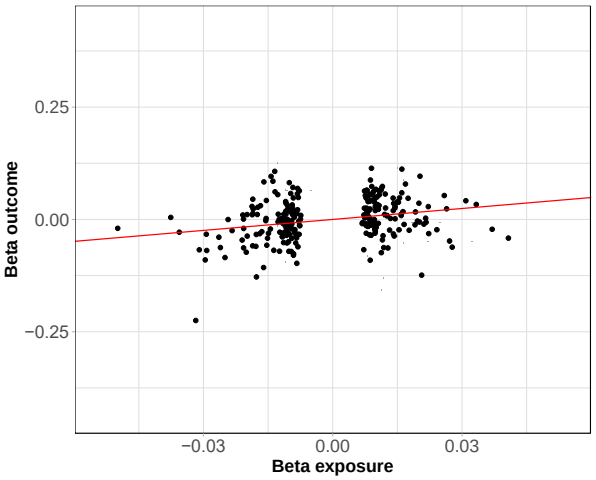




Phecode: 414; Other forms of chronic heart disease

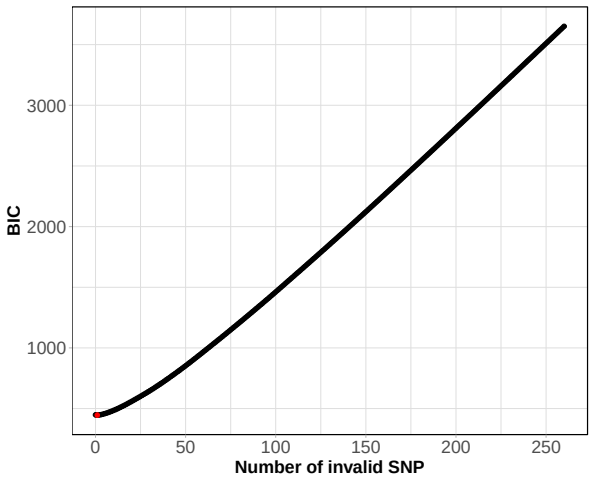
N invalid IV = 0

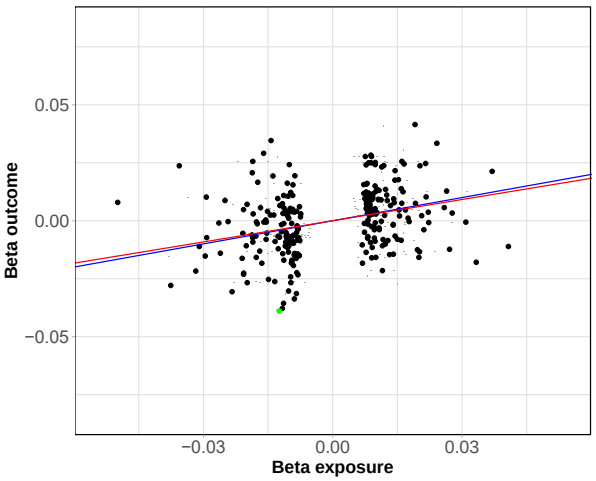




Phecode: 418; Nonspecific chest pain

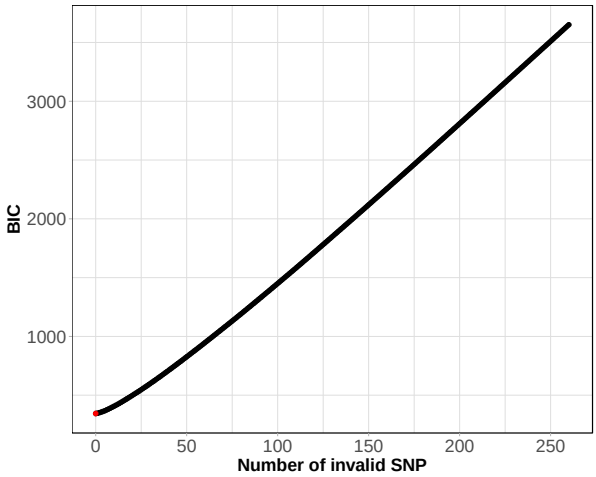
N invalid IV = 1

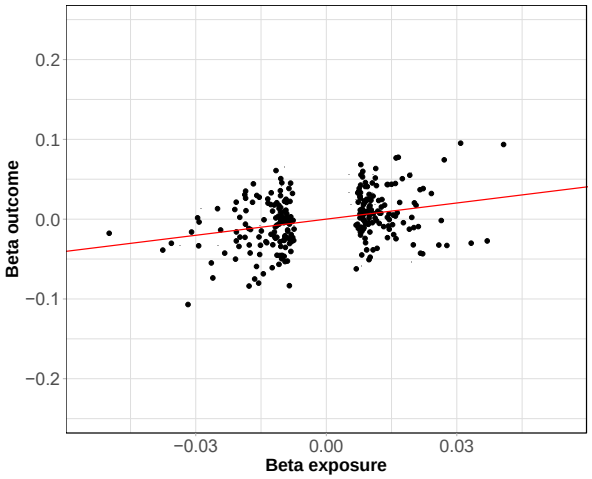




Phecode: 428.2; Heart failure NOS

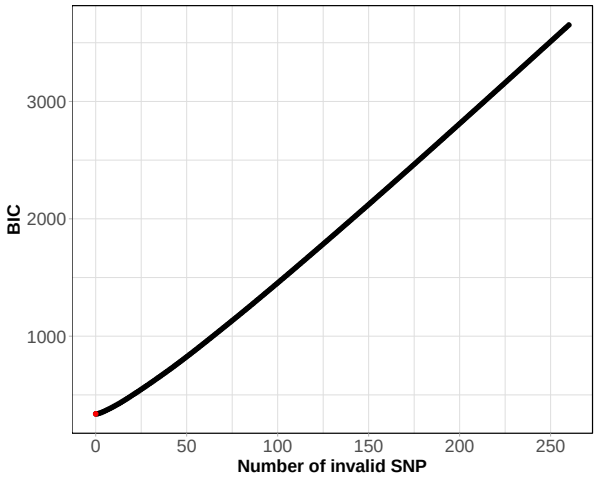
N invalid IV = 0

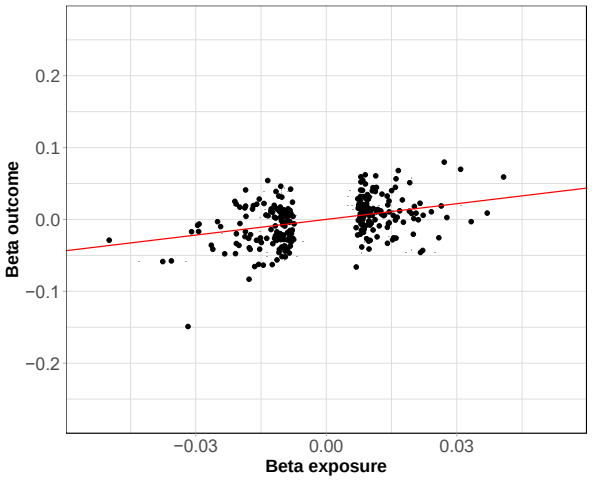




Phecode: 428; Congestive heart failure; nonhyperte

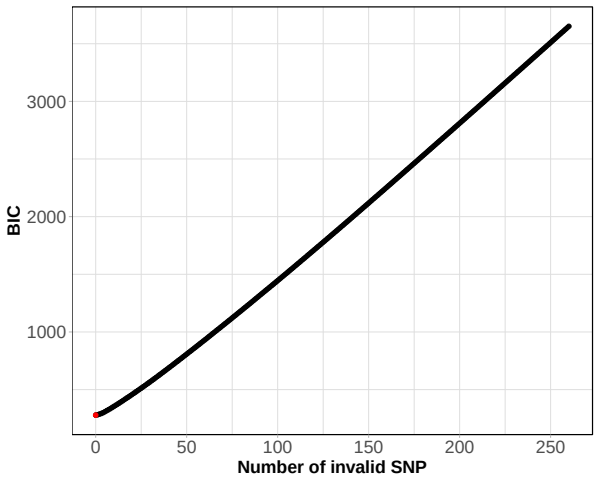
N invalid IV = 0

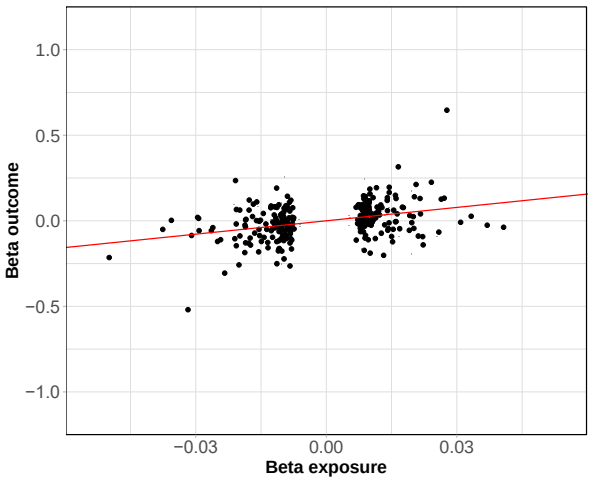




Phecode: 443.7; Peripheral angiopathy in diseases c

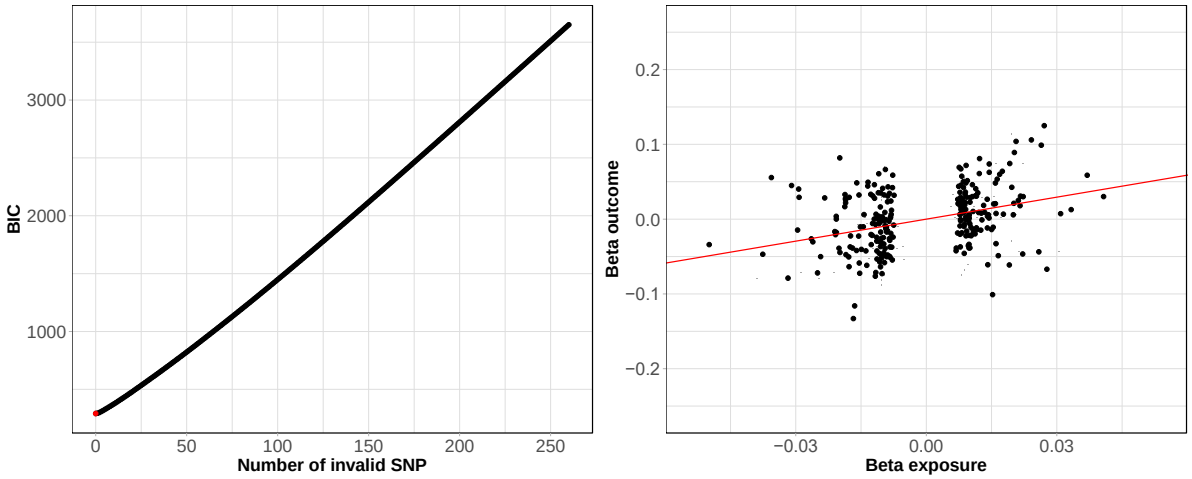
N invalid IV = 0





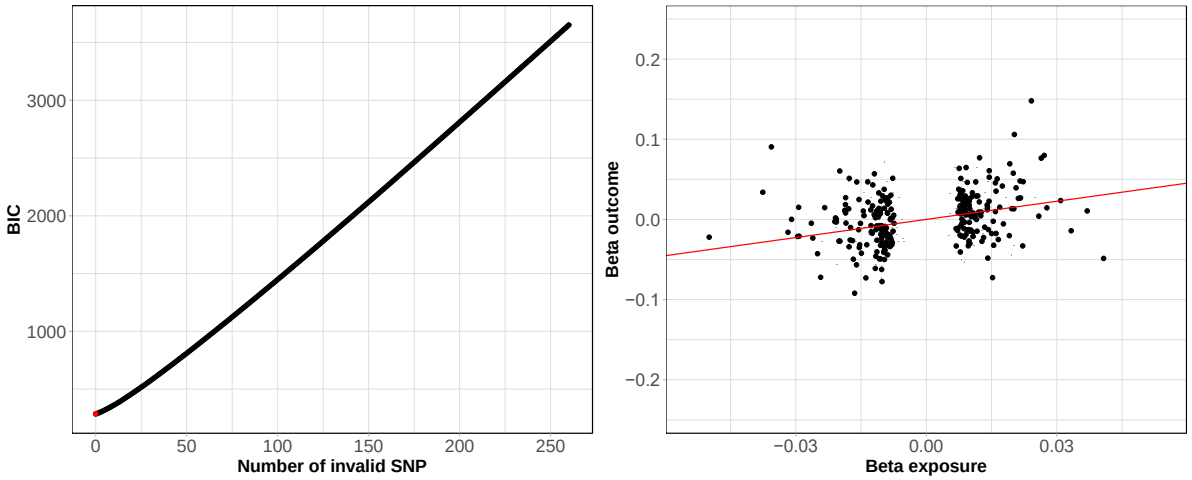
Phecode: 443.9; Peripheral vascular disease, unspec

N invalid IV = 0



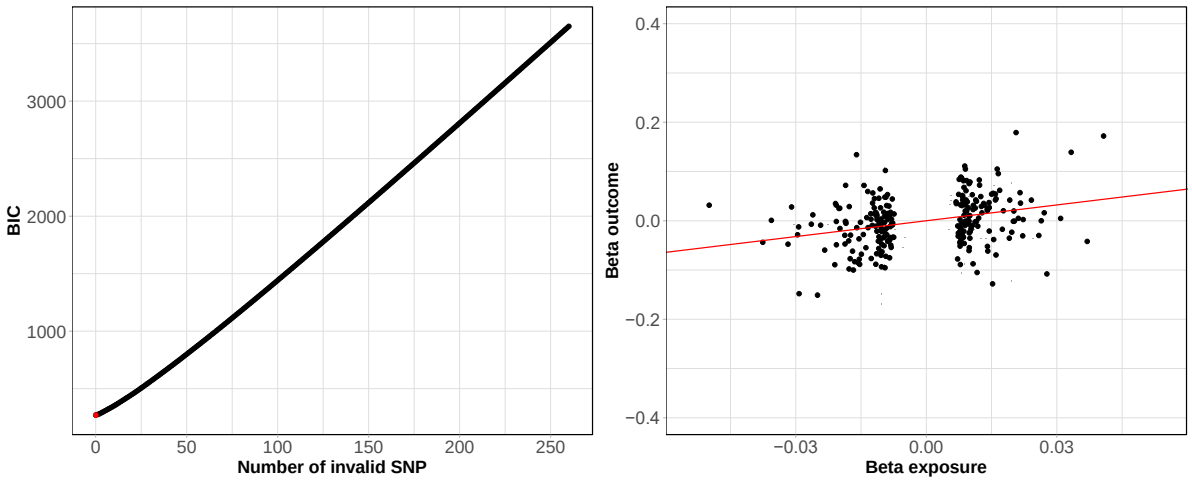
Phecode: 443; Peripheral vascular disease

N invalid IV = 0



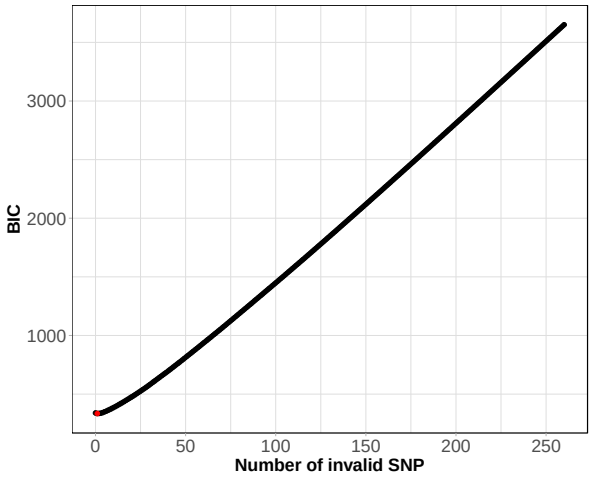
Phecode: 447; Other disorders of arteries and arteri

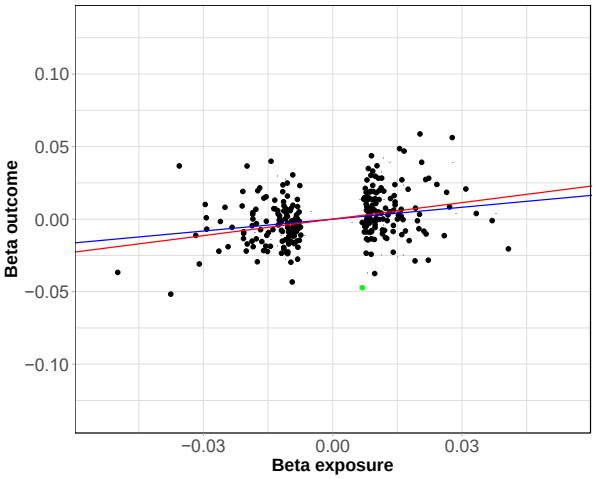
N invalid IV = 0



Phecode: 459.9; Circulatory disease NEC

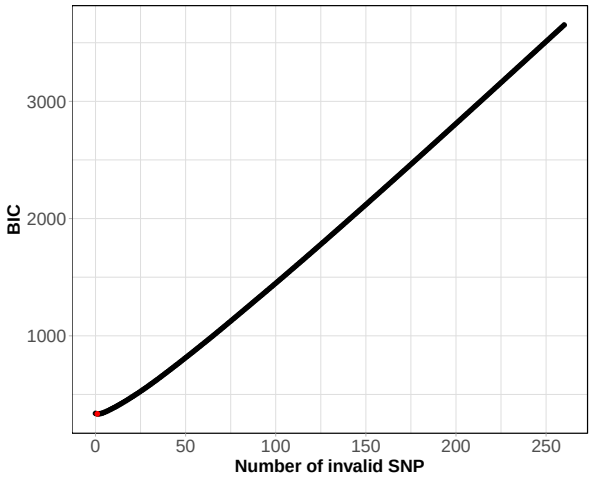
N invalid IV = 1

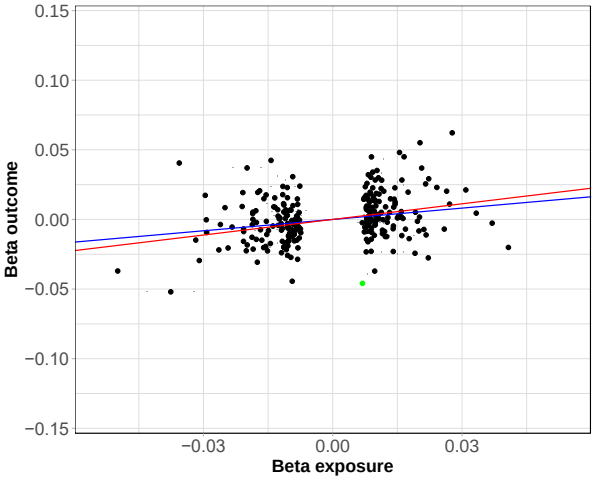




Phecode: 459; Other disorders of circulatory systen

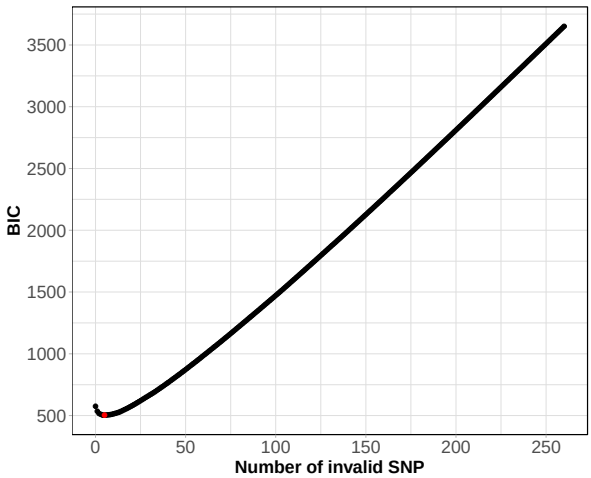
N invalid IV = 1

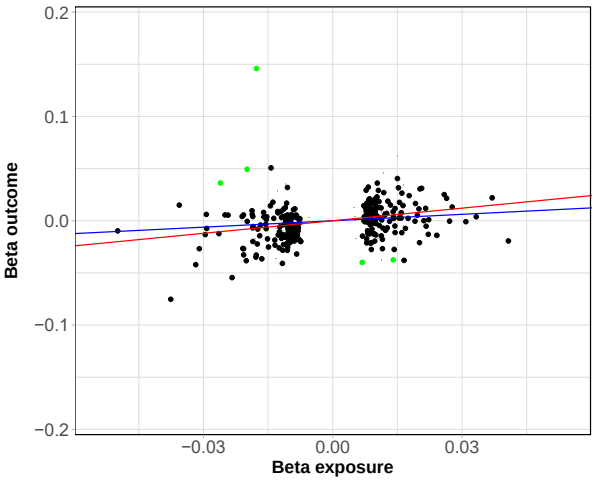




Phecode: 495; Asthma

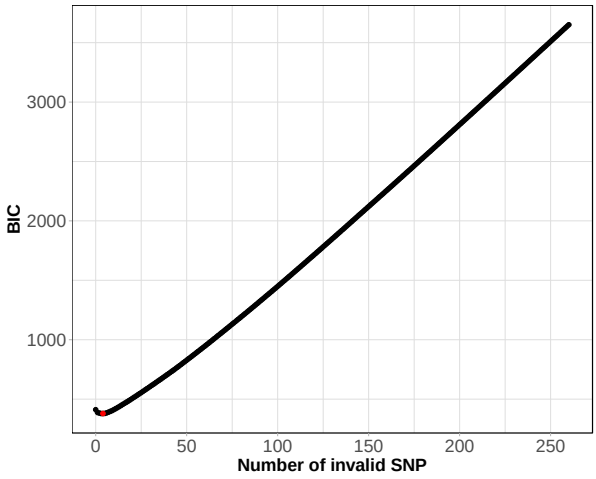
N invalid IV = 5

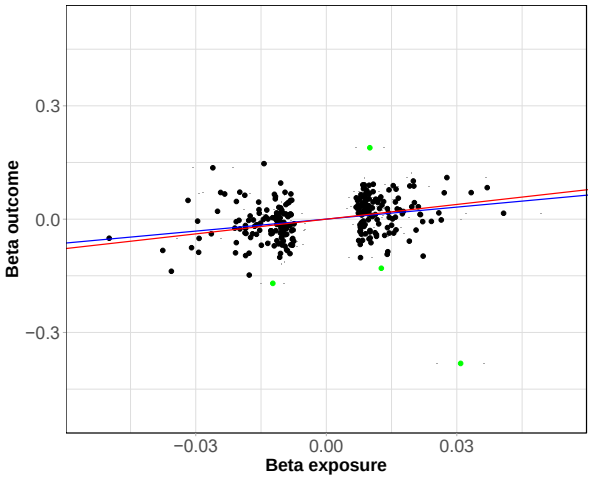




Phecode: 571.5; Other chronic nonalcoholic liver dis

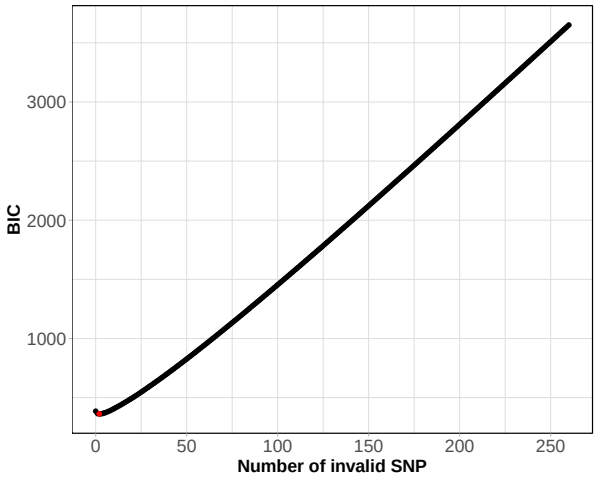
N invalid IV = 4

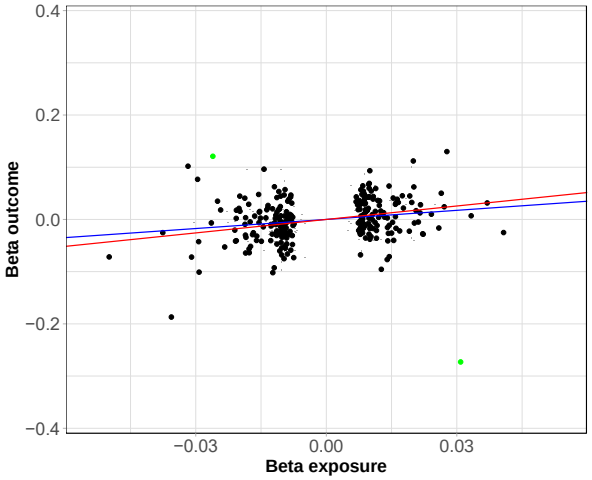




Phecode: 571; Chronic liver disease and cirrhosis

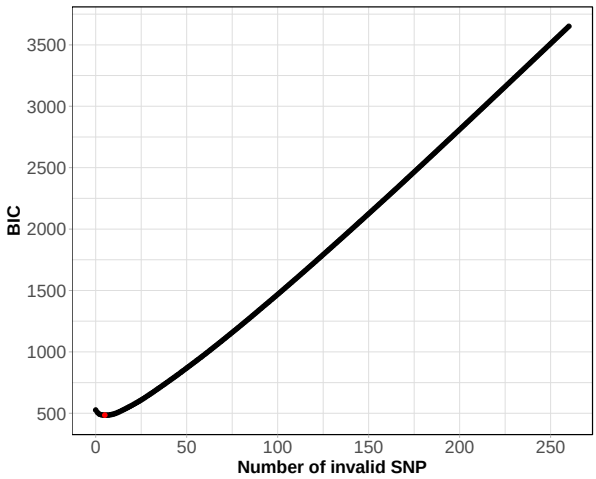
N invalid IV = 2

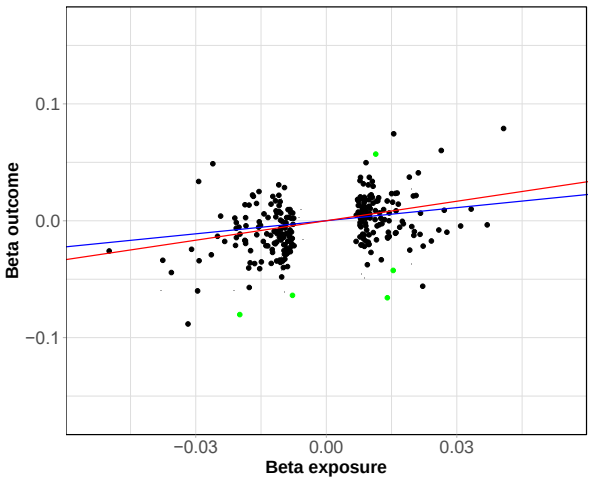




Phecode: 574.1; Cholelithiasis

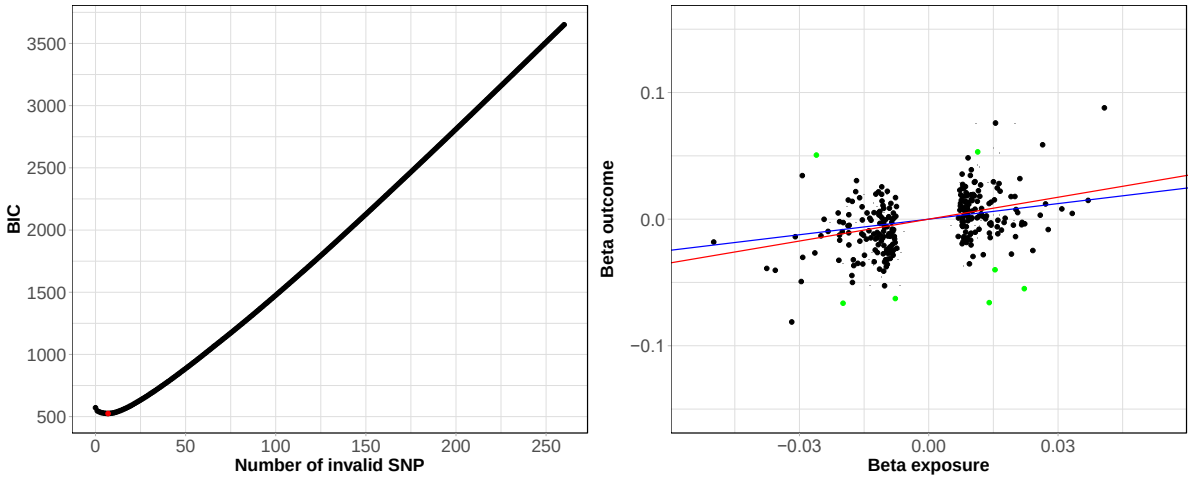
N invalid IV = 5





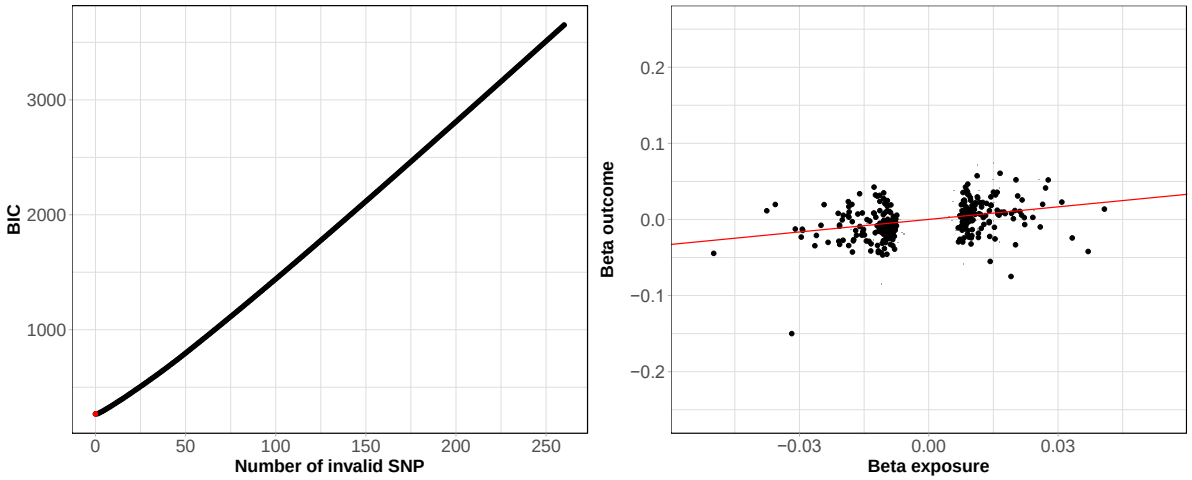
Phecode: 574; Cholelithiasis and cholecystitis

N invalid IV = 7



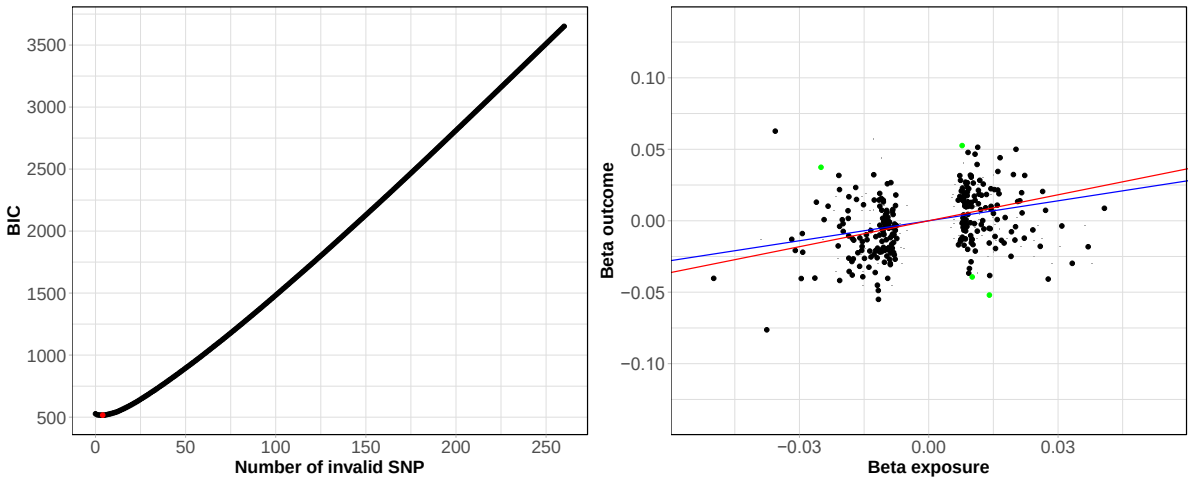
Phecode: 585; Renal failure

N invalid IV = 0



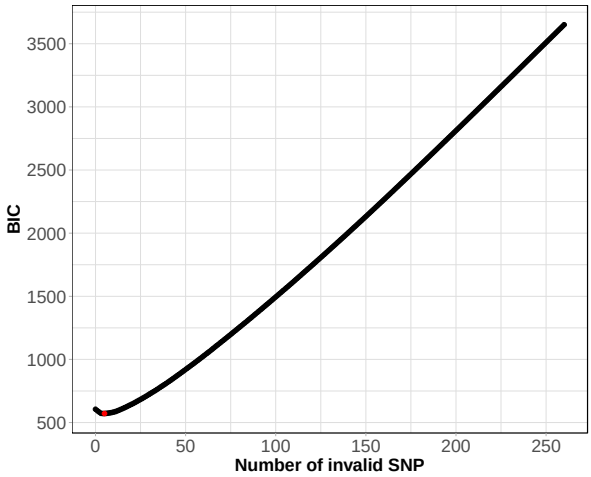
Phecode: 716.2; Unspecified monoarthritis

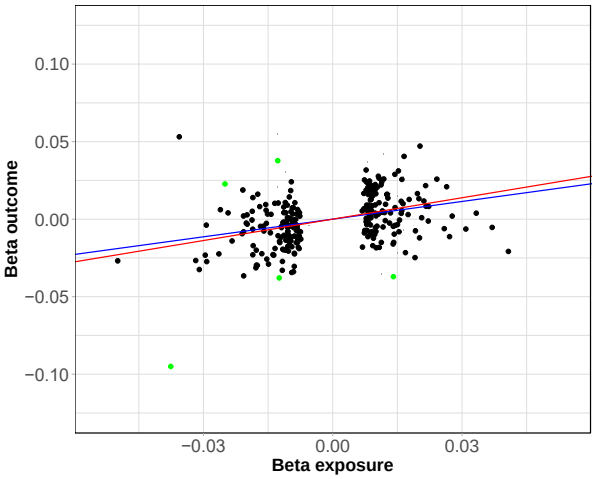
N invalid IV = 4



Phecode: 716.9; Arthropathy NOS

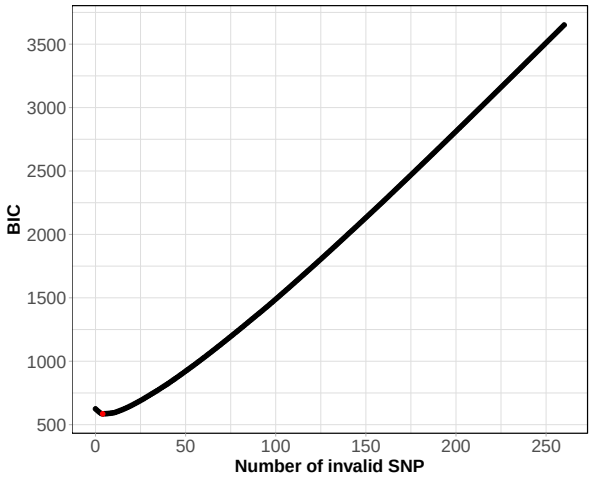
N invalid IV = 5

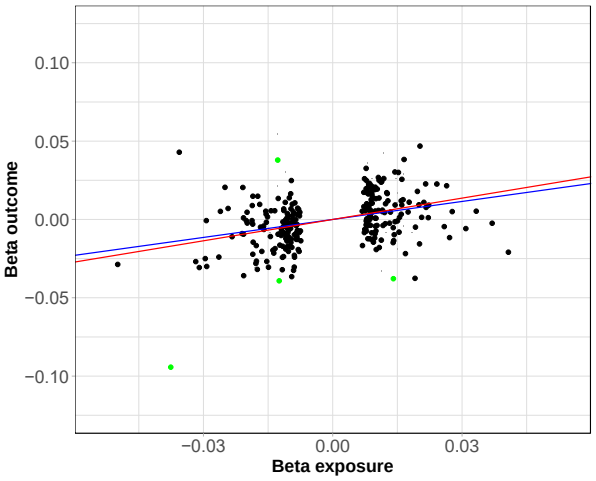




Phecode: 716; Other arthropathies

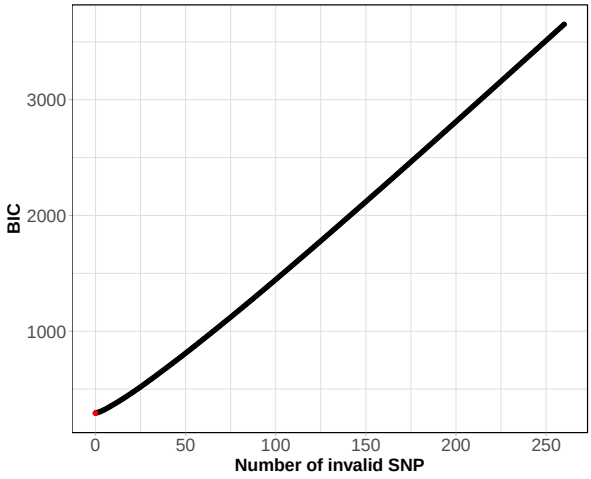
N invalid IV = 4

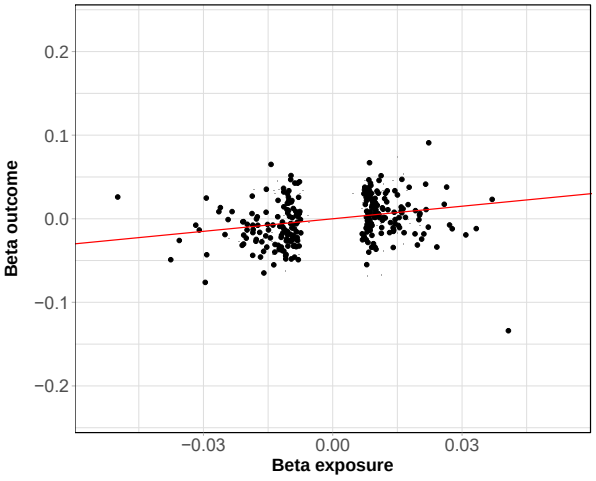




Phecode: 721.1; Spondylosis without myelopathy

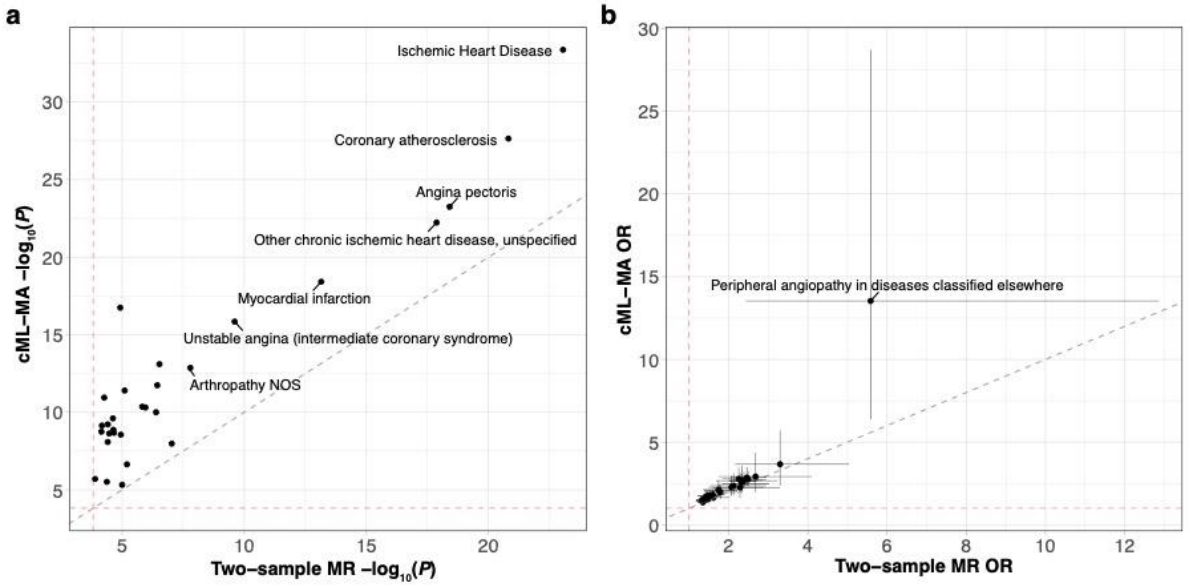
N invalid IV = 0



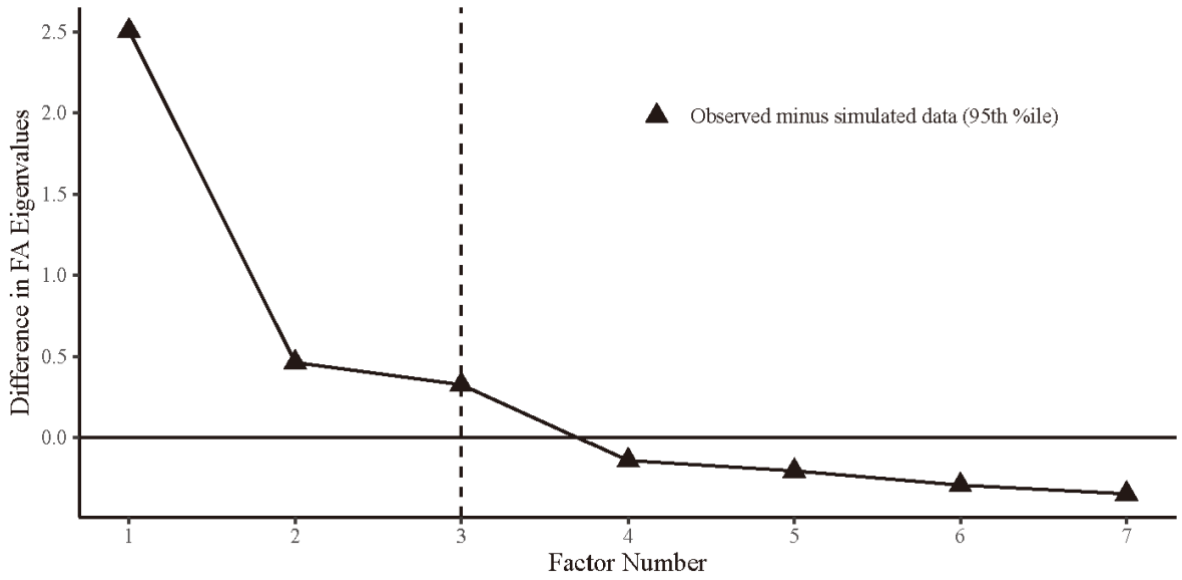


Supplementary Figure 14. Scatter plot of comparison between two-sample Mendelian randomization (TSMR) and constrained maximum likelihood and model averaging-based Mendelian randomization (cML-MA MR) results for 29 health outcomes

a, Scatter plot of uncorrected -$\log_{10}(P)$ compared with x-axis from TSMR and the y-axis from cML-MA MR. The *P* for TSMR depends on the method used (e.g., two-sided *z*-test for inverse variance weighted and two-sided *t*-test for MR-PRESSO). The *P* for cML-MA MR is from two-sided *z*-test. The red dashed line represents the Bonferroni correction threshold of 0.05/29, and the gray dashed line represents the identity line. **b**, Scatter plot of odds ratio (OR) comparison with *x*-axis from TSMR and *y*-axis from cML-MA MR. The error bars represent 95% confidence intervals, calculated as OR ± 1.96 × s.e. The red dashed line represents OR = 1, and the gray dashed line represents the identity line. The sample size of health outcomes is available in **Supplementary Table 49**.



Supplementary Figure 15. Scree plot from parallel analysis for UKB-excluded cohorts
The scree plot was obtained by leveraging the genetic correlation matrix of seven MetS components from LD score regression. The *x*-axis represents the number of factors, and the *y*-axis represents the difference between the eigenvalues computed from the LDSC-derived genetic correlation matrix and the Monte-Carlo simulated genetic correlation matrix. The dotted black line indicates the suggested number of factors to retain for exploratory factor analysis.



Supplementary References

1. Alberti, K.G. *et al.* Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; American heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity. *Circulation* **120**, 1640-1645 (2009).
2. Grundy, S.M., Brewer, H.B., Cleeman, J.I., Smith, S.C. & Lenfant, C. Definition of Metabolic Syndrome. *Circulation* **109**, 433-438 (2004).
3. Eckel, R.H., Grundy, S.M. & Zimmet, P.Z. The metabolic syndrome. *The Lancet* **365**, 1415-1428 (2005).
4. Alberti, K.G.M.M. & Zimmet, P.Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation. *Diabetic Medicine* **15**, 539-553 (1998).
5. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Final Report. *Circulation* **106**, 3143-3143 (2002).
6. Einhorn, D. American College of Endocrinology Position Statement on the Insulin Resistance Syndrome *. *Endocrine Practice* **9**, 5-21 (2003).
7. Alberti, K.G.M.M., Zimmet, P. & Shaw, J. The metabolic syndrome–a new worldwide definition. *The Lancet* **366**, 1059-1062 (2005).
8. Grundy, S.M. *et al.* Diagnosis and Management of the Metabolic Syndrome. *Circulation* **112**, 2735-2752 (2005).
9. Maier, R. *et al.* Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. *The American Journal of Human Genetics* **96**, 283-294 (2015).
10. Maier, R.M. *et al.* Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications* **9**, 989 (2018).
11. Li, C., Yang, C., Gelernter, J. & Zhao, H. Improving genetic risk prediction by leveraging pleiotropy. *Human Genetics* **133**, 639-650 (2014).
12. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164-e164 (2010).
13. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
14. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310-315 (2014).
15. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22**, 1790-1797 (2012).
16. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-

330 (2015).
17. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215-216 (2012).
18. Jansen, P.R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature Genetics* **51**, 394-403 (2019).
19. Ratnapriya, R. *et al.* Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nature Genetics* **51**, 606-610 (2019).
20. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics* **53**, 1290-1299 (2021).
21. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
22. van der Wijst, M.G.P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* **50**, 493-497 (2018).
23. Schmiedel, B.J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715.e16 (2018).
24. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* **53**, 1300-1310 (2021).
25. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics* **45**, 1238-1243 (2013).
26. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics* **44**, 1084-1089 (2012).
27. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature Neuroscience* **20**, 1418-1426 (2017).
28. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience* **19**, 1442-1453 (2016).
29. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neuroscience* **17**, 1418-1428 (2014).
30. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580-585 (2013).
31. Schmitt, Anthony D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* **17**, 2042-2059 (2016).
32. Giusti-Rodríguez, P. *et al.* Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. *bioRxiv*, 406330 (2019).
33. Forrest, A.R.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-470 (2014).
34. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).

35. Karlsson Linnér, R. *et al.* Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. *Nature Neuroscience* **24**, 1367-1376 (2021).
36. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228-1235 (2015).
37. Fehrmann, R.S.N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics* **47**, 115-125 (2015).
38. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740 (2011).
39. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics* **54**, 1.30.1-1.30.33 (2016).
40. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research* **51**, D1353-D1359 (2023).
41. Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Research* **51**, D1038-D1045 (2023).
42. Kraja, A.T. *et al.* Pleiotropic genes for metabolic syndrome and inflammation. *Molecular Genetics and Metabolism* **112**, 317-338 (2014).
43. Zhang, G., Lv, X., Yang, Q. & Liu, H. Identification of HM13 as a prognostic indicator and a predictive biomarker for immunotherapy in hepatocellular carcinoma. *BMC Cancer* **22**, 888 (2022).
44. Napoli, C., Schiano, C. & Soricelli, A. Increasing evidence of pathogenic role of the Mediator (MED) complex in the development of cardiovascular diseases. *Biochimie* **165**, 1-8 (2019).
45. Solomon, S.S., Majumdar, G., Martinez-Hernandez, A. & Raghow, R. A critical role of Sp1 transcription factor in regulating gene expression in response to insulin and other hormones. *Life Sciences* **83**, 305-312 (2008).
46. Ow, J.R., Tan, Y.H., Jin, Y., Bahirvani, A.G. & Taneja, R. Chapter Nine - Stra13 and Sharp-1, the Non-Grouchy Regulators of Development and Disease. in *Current Topics in Developmental Biology*, Vol. 110 (ed. Taneja, R.) 317-338 (Academic Press, 2014).
47. Kim, Y., Han, B.-G. & the Ko, G.E.S.g. Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium. *International Journal of Epidemiology* **46**, e20-e20 (2017).
48. Moon, S. *et al.* The Korea Biobank Array: Design and Identification of Coding Variants Associated with Blood Biochemical Traits. *Scientific Reports* **9**, 1382 (2019).
49. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284-1287 (2016).
50. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48**, 1443-1448 (2016).
51. Choi, S.W. & O'Reilly, P.F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, giz082 (2019).

52. Márquez-Luna, C. *et al.* Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nature Communications* **12**, 6052 (2021).
53. Yun, J.-S. *et al.* Polygenic risk for type 2 diabetes, lifestyle, metabolic health, and cardiovascular disease: a prospective UK Biobank study. *Cardiovascular Diabetology* **21**, 131 (2022).
54. Xue, H., Shen, X. & Pan, W. Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal of Human Genetics* **108**, 1251-1269 (2021).